Automatica 151 (2023) 110912

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Separation of learning and control for cyber–physical systems☆

Andreas A. Malikopoulos

Department of Mechanical Engineering, University of Delaware, 130 Academy Street, Newark, DE, 19716, USA

ARTICLE INFO

Article history:
Received 24 May 2022
Received in revised form 18 September 2022
Accepted 19 December 2022
Available online xxxx

Keywords:
Separation of learning and control
Stochastic optimal control
Information state
Markov decision theory

ABSTRACT

Most cyber–physical systems (CPS) encounter a large volume of data which is added to the system gradually in real time and not altogether in advance. In this paper, we provide a theoretical framework that yields optimal control strategies for such CPS at the intersection of control theory and learning. In the proposed framework, we use the actual CPS, i.e., the "true" system that we seek to optimally control online, in parallel with a model of the CPS that is available. We then institute an information state for the system which does not depend on the control strategy. An important consequence of this independence is that for any given choice of a control strategy and a realization of the system's variables until time t, the information states at future times do not depend on the choice of the control strategy at time t but only on the realization of the decision at time t, and thus they are related to the concept of separation between estimation of the state and control. Namely, the future information states are separated from the choice of the current control strategy. Such control strategies are called separated control strategies. Hence, we can derive offline the optimal control strategy of the system with respect to the information state, which might not be precisely known due to model uncertainties or complexity of the system, and then use standard learning approaches to learn the information state online while data are added gradually to the system in real time. We show that after the information state becomes known, the separated control strategy of the CPS model derived offline is optimal for the actual system. We illustrate the proposed framework in a dynamic system consisting of two subsystems with a delayed sharing information structure.

1. Introduction

1.1. Motivation

Cyber–physical systems (CPS), in many instances, represent systems of subsystems with an informationally decentralized structure such as networked control systems, emerging mobility systems, communication networks, digital twin, and internet of things. Systems with informationally decentralized structures impose significant challenges compared to systems with centralized *information structures*; see van Schuppen and Villa (2015). The information structure in a system designates what information each subsystem knows about the status of the system and when. Several efforts on the characterization of information structures and their implications on optimality results have been reported in the literature over the years; see Mahajan et al. (2012), Subramanian et al. (2022), Witsenhausen (1971). The information structure in a system stipulates the complexity, i.e., see Papadimitriou and Tsitsiklis (1982), Tsitsiklis and Athans (1985), of the optimal control problem and can lead to computational implications; see Papadimitriou and Tsitsiklis (1985). The latter depends on whether the system has a *strictly classical* information structure or a *nonclassical* information structure. In classical information structures, all subsystems receive the same information and have perfect recall; see Malikopoulos (2016). If there is only one subsystem, then such information structures are called *strictly classical* resulting in typical centralized stochastic control problems; see Kumar and Varaiya (1986), Kushner (1971). In partially nested information structures, there are some subsystems who have a nonempty intersection of their information structures while they have perfect recall. Any information structure that is not classical, or partially nested, is called nonclassical.

In most CPS applications with nonclassical information structures there is a large volume of data of a dynamic nature which is added to the system gradually in real time and not altogether in advance. As the volume of data increases, the domain of the control strategies also increases, and thus it becomes challenging to search for an optimal strategy. Even if an optimal strategy is found, implementing such strategies with increasing domains is burdensome. In such applications, we typically assume an ideal model of the system to derive optimal control strategies.

☆ This work was supported by National Science Foundation, USA under Grants CNS-2149520 and CMMI-2219761. The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Kyriakos G. Vamvoudakis under the direction of Editor Miroslav Krstic.

E-mail address: andreas@udel.edu.

https://doi.org/10.1016/j.automatica.2023.110912
0005-1098/

Such model-based control approaches cannot effectively facilitate optimal solutions with performance guarantees due to the discrepancy between the model and the actual CPS. On the other hand, traditional supervised learning approaches cannot always facilitate robust solutions using data derived offline. By contrast, applying reinforcement learning approaches directly to the actual CPS might impose significant implications on safety and robust operation of the system.

The goal of this paper is to provide a theoretical framework that aims at separating the control and learning tasks which allows us to combine offline model-based control with online learning approaches, and thus circumvent the challenges in deriving optimal strategies for CPS with nonclassical information structures. The framework can fit well in applications related to digital twins where a virtual representation of a real-world physical system serves as the indistinguishable digital counterpart of it.

1.2. Related work

1.2.1. Model-based control

Most CPS represent systems of subsystems with nonclassical information structures imposing the following technical challenges (Papadimitriou & Tsitsiklis, 1987): (a) the functional optimization problem of selecting the optimal strategy is not trivial as the class of strategies is infinite dimensional, and (b) the data increase with time causing significant implications on storage requirements and real-time implementation. These difficulties can be addressed by finding sufficient statistics to compress the growing data without loss of optimality (Striebel, 1965) using a conditional probability of the state of the system at time t given all the data available up until time t. This conditional probability is called *information state*, and it takes values in a time-invariant space. This information state can help us derive results for optimal control strategies in a time-invariant domain; Krishnamurthy (2016).

One key property of such information states is that they do not depend on the control strategy of the system, and thus they are related to the concept of separation between estimation and control. An important consequence of this separation is that for any given choice of control strategies and a realization of the system's variables until time t, the information states at future times do not depend on the choice of the control strategy at time t but only on the realization of the decision at time t; see Malikopoulos (2023). Thus, the future information states are separated from the choice of the current control strategy. The latter is necessary in order to formulate a classical dynamic program (Bertsekas, 2017; Howard, 1960), where at each step the optimization problem is to find the optimal decision for a given realization of the information state.

Several optimality results using information states defined in time-invariant spaces have been reported in the literature for systems with nonclassical information structures; see Dave and Malikopoulos (2019, 2020), Gupta et al. (2015), Kurtaran (1979), Nayyar et al. (2011), Varaiya and Walrand (1978), Witsenhausen (1971), Wu and Lall (2014). There are three main approaches to address optimal control problems with a nonclassical information structure: (1) the *person-by-person* approach, (2) the *designer's* approach, and (3) the *common information* approach. The person-by-person approach (McGuire & Radner, 1972) aims to convert the problem into a centralized stochastic control problem from the point of view of each subsystem. Namely, we arbitrarily fix the strategies for all subsystems except for one, say subsystem $k \in \mathcal{K}$, $\mathcal{K} = \{1, \ldots, K\}$, $K \in \mathbb{N}$, and then, we derive the optimal strategy for k given the strategies for all other subsystems. We repeat this process for all subsystems until no subsystem can improve the performance of the system by unilaterally changing their strategy. The designer's approach was first introduced by Witsenhausen (1973), as a standard form for sequential stochastic control with a nonclassical information structure, and extended later by Mahajan (2008). The designer's approach transforms the problem into a centralized, open-loop planning problem where the objective is to derive the optimal control strategy of the system before the system starts evolving. Thus, no data are observed by the designer, and thus this approach leads to a dynamic programming decomposition over a space of functions instead of decisions imposing significant computational implications; see Papadimitriou and Tsitsiklis (1987). Finally, in the common information approach (Nayyar et al., 2011, 2013), the subsystems share a subset of their past observations and decisions to a shared memory accessible by all subsystems. The solution is derived by reformulating the problem from the viewpoint of a "coordinator" with access only to the shared information (the common information), whose task is to provide "prescription" strategies to each subsystem. The coordinator's problem is a centralized stochastic control problem.

1.2.2. Learning-based control

Adaptive control methods (Åström & Wittenmark, 1995; Ioannou & Sun, 1996; Narendra & Annaswamy, 1989; Sastry & Bodson, 1989) have successfully addressed regulation and tracking control problems with safety guarantees by accommodating model uncertainties; see Dydek et al. (2013), Leman et al. (2009). Reinforcement learning (RL) has emerged from machine learning as an adaptive approach to control dynamical systems; Bertsekas and Tsitsiklis (1996), Sutton and Barto (1998). Several efforts have focused on safe learning approaches combining robust reachability guarantees from control theory with Bayesian analysis based on empirical observations (Fisac et al., 2019), and on learning the system's unknown dynamics based on a Gaussian process model to iteratively approximate the maximal safe set; see Akametalu et al. (2014). Iterative learning control (Armstrong et al., 2021), has been also widely used for system identification, or in conjunction with extremum seeking (Khong et al., 2016a, 2016b), for recursively constructing an input such that the corresponding system output tracks a prescribed reference trajectory closely. In communication networks, where models of wireless channels are available only through data samples (Gatsis & Pappas, 2021) there have been efforts on learning approximately optimal power allocation policies to maximize control performance of a set of independent control systems within a fixed budget; see Eisen et al. (2018).

Other research efforts over the years have focused on developing robust learning-based approaches in applications related to quadrotor safety and steady-state stability (Aswani et al., 2013), learning-based model predictive control (Rosolia & Borrelli, 2018), real-time learning (Malikopoulos, 2009) of powertrain operation of vehicles with respect to the driver's driving style (Malikopoulos et al., 2010), learning for traffic control in simulation (Wu et al., 2017) in conjunction with transfer of learned policies from simulation to a scaled environment (Chalaki et al., 2020), decentralized learning for stochastic games (Arslan & Yüksel, 2017), learning for optimal social routing (Krichene et al., 2018) and congestion games (Krichene et al., 2015), and learning for enhanced security against replay attacks in CPS; see Sahoo and Vamvoudakis (2020), Zhai and Vamvoudakis (2021).

Regularities of optimal control on the space of transition kernels along with the implications on robustness of optimal control strategies derived using an "incorrect" model and applied to the actual system have been discussed by Kara and Yüksel (2018). Approximate planning and learning in partially observed systems using an information state was more recently proposed

by Subramanian et al. (2022). Alternatively, one can establish an approximate information state, defined in terms of properties that can be estimated using sampled trajectories, along with an approximate dynamic program; see Subramanian and Mahajan (2019). This approach provides a constructive way for RL in partially observed systems. Other efforts have also combined model reference adaptive control with RL to generate online policies; see Guha and Annaswamy (2021). Two recent survey papers by Kiumarsi et al. (2018) and Recht (2019) provide a comprehensive review of the general RL problem formulations along with a complete list of applications.

1.3. Contributions of this paper

In this paper, we consider CPS consisting of several subsystems with a common objective and a nonclassical information structure, where the state of the system is not fully observed. We provide a theoretical framework, which can combine offline model-based control with online learning approaches, to yield the optimal control strategy of the system. More specifically, we identify a sufficient information state for the system which does not depend on the control strategy. An important consequence of this independence is that for any given choice of a control strategy and a realization of the system's variables until time t, the information states at future times do not depend on the choice of the control strategy at time t but only on the realization of the decision at time t, and thus they are related to the concept of separation between estimation of the state and control. Namely, the future information states are separated from the choice of the current control strategy. The adjective "separated" is used to emphasize the fact that in implementing such an optimal policy, we first need to learn the information state and then choose the control. Such control strategies are called separated control strategies. Hence, we can derive offline the optimal control strategy of the system with respect to the information state, which might not be precisely known due to model uncertainties or complexity of the system, and then use standard learning approaches to learn the information state online while data are added gradually to the system in real time.

The contributions of this paper are: (1) the institution of an information state of the system, which does not depend on the control strategy (Theorem 1), that allows us to restrict attention to separated control strategies; (2) a dynamic programming decomposition that uses a CPS model and the information state to derive offline optimal separated control strategies (Theorem 2) which are optimal for the actual system (Theorem 3); and (3) providing structural properties of the dynamic programming decomposition (Theorem 4) which allow us to derive the optimal strategies offline using standard techniques for centralized partially observed Markov decision processes.

The two features which sharply distinguish the framework presented here from previous learning-based, or combined learning and control approaches reported in the literature to date are the following. First, the CPS imposes a nonclassical information structure while the state of the system is not fully observed. To the best of our knowledge, this is the first time that results on such systems are derived by separating the control and the learning tasks of the problem. Second, the large volume of data that is added to the system gradually is compressed to sufficient statistics without loss of optimality (Theorem 2) which constitutes the information state of the system. Using this information state, we derive results for optimal control strategies in a time-invariant domain. Thus, the volume of data which is added gradually to the system does not cause the domain of the control strategies to increase with time. The latter is quite important since searching and then implementing control strategies with increasing domain is burdensome.

1.4. Organization of this paper

The remainder of the paper proceeds as follows. In Section 2, we provide the modeling framework and the formulation of the optimal control problem for a CPS with nonclassical information structure. In Section 3, we present the analysis for deriving separated control strategies. In Section 4, we present a simple example to illustrate the proposed framework. Finally, we provide concluding remarks and discuss potential directions for future research in Section 5.

2. Problem formulation

2.1. Notation

Subscripts denote time, and superscripts index subsystems. We denote random variables with upper case letters, and their realizations with lower case letters, e.g., for a random variable X_t, x_t denotes its realization. The shorthand notation $X_t^{1:K}$ denotes the vector of random variables $\left(X_t^1, X_t^2, \ldots, X_t^K\right)$, $x_t^{1:K}$ denotes the vector of their realization $\left(x_t^1, x_t^2, \ldots, x_t^K\right)$, and $h_t^{1:K}(\cdot,\cdot)$ denotes the vector of functions $\left(h_t^1(\cdot,\cdot), \ldots, h_t^K(\cdot,\cdot)\right)$. The expectation of a random variable is denoted by $\mathbb{E}[\cdot]$, the probability of an event is denoted by $\mathbb{P}(\cdot)$, and the probability density function is denoted by $p(\cdot)$. For a control strategy $\mathbf{g}$, we use $\mathbb{E}^{\mathbf{g}}[\cdot]$, $\mathbb{P}^{\mathbf{g}}(\cdot)$, and $p^{\mathbf{g}}(\cdot)$ to denote that the expectation, probability, and probability density function, respectively, depend on the choice of the control strategy $\mathbf{g}$. For two measurable spaces $(\mathcal{X}, \mathscr{X})$ and $(\mathcal{Y}, \mathscr{Y})$, $\mathscr{X} \otimes \mathscr{Y}$ is the product σ-algebra on $\mathcal{X} \times \mathcal{Y}$ generated by the collection of all measurable rectangles, i.e., $\mathscr{X} \otimes \mathscr{Y} := \sigma(\{A \times B : A \in \mathscr{X}, B \in \mathscr{Y}\})$. The product of $(\mathcal{X}, \mathscr{X})$ and $(\mathcal{Y}, \mathscr{Y})$ is the measurable space $(\mathcal{X} \times \mathcal{Y}, \mathscr{X} \otimes \mathscr{Y})$. We denote the Cartesian product of the sets $\mathcal{G}^k$, $k \in \mathcal{K}$, $\mathcal{K} = \{1, \ldots, K\}$, $K \in \mathbb{N}$, with $\times_{k\in\mathcal{K}}\mathcal{G}^k$.

2.2. Proposed approach

We consider a CPS representing a system of subsystems with an informationally decentralized structure in which there is a large volume of data of a dynamic nature that is added to the system gradually and not altogether in advance. For such systems, using model-based control approaches cannot effectively facilitate optimal solutions with performance guarantees due to the discrepancy between the model and the actual CPS. On the other hand, since there is a large volume of data of a dynamic nature that is added to the system gradually in real time, traditional supervised learning approaches might not facilitate robust solutions using data derived offline. By contrast, applying reinforcement learning approaches directly to the actual CPS might impose significant implications on safety and robust operation of the system.

To address these challenges, our framework aims at separating the control and learning tasks which eventually allows us to combine offline model-based control with online learning approaches. In particular, we aim at identifying a sufficient information state for the CPS that takes values in a time-invariant space, and use this information state to derive separated control strategies. Separated control strategies are related to the concept of separation between the estimation of the information state and control of the system. An important consequence of this separation is that for any given choice of control strategies and a realization of the system's variables until time t, the information states of the system at future times do not depend on the choice of the control strategy at time t but only on the realization of the control at time t; see Kumar and Varaiya (1986). Thus, the future information states are separated from the choice of the current control strategy. By establishing separated control strategies, we

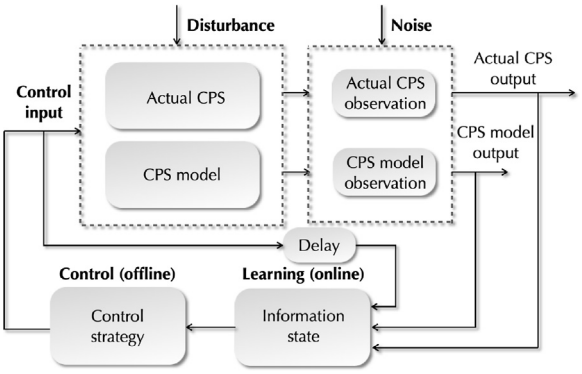

Fig. 1. Illustration of the proposed framework.

can derive offline the optimal control strategy of the system with respect to the information state, which might not be precisely known due to model uncertainties or complexity of the system, and then use learning methods to learn the information state online while data are added gradually to the system in real time.

More specifically, in the proposed framework illustrated in Fig. 1, we use the actual CPS, i.e., the actual system that we seek to optimally control online, in parallel with a model of the CPS that is available. The main idea here is the institution of an information state which is the conditional joint probability distribution of the states of the CPS model and the actual CPS at time t given all data available of the model up until time t, i.e., p(state of CPS model, state of actual CPS | data of the CPS model). We use this information state along with the CPS model to derive offline separated control strategies. Since we derive the optimal strategies offline, the state of the actual CPS is not known, i.e., the actual CPS operates only online, and thus the optimal strategy of the CPS model is parameterized with respect to all realizations of the state of the actual CPS. However, the control strategy and the process of estimating the information state are separated. Therefore, we can learn the information state of the system online, while we operate simultaneously the CPS model and the actual CPS in real time. Namely, the optimal strategy derived for the CPS model offline, which is parameterized with respect to the state of the actual CPS, is used to operate the actual CPS in parallel with the CPS model. As we collect data from the two systems, we can learn the information state online. In our exposition, we show that when the information state becomes known online through learning, the separated control strategy of the CPS model derived offline is optimal for the actual CPS (Theorem 3). The framework described above is centralized, e.g., a central controller controls all subsystems.

2.3. Modeling framework

We consider a CPS consisting of $K \in \mathbb{N}$ subsystems with a measurable state space $(\mathcal{X}_t, \mathscr{X}_t)$, where $\mathcal{X}_t$ is the set in which the CPS state takes values at time $t = 0, 1, \ldots, T$, $T \in \mathbb{N}$, and $\mathscr{X}_t$ is the associated σ-algebra. Let X_t be a random variable that represents the state of the CPS model and $\hat{X}_t$ be a random variable that represents the state of the actual CPS. Both random variables are defined on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$, i.e., $X_t : (\Omega, \mathscr{F}) \to (\mathcal{X}_t, \mathscr{X}_t)$, $\hat{X}_t : (\Omega, \mathscr{F}) \to (\mathcal{X}_t, \mathscr{X}_t)$, where Ω is the sample space, $\mathscr{F}$ is the associated σ-algebra, and $\mathbb{P}$ is a probability measure on $(\Omega, \mathscr{F})$. The control of each subsystem $k \in \mathcal{K}$, $\mathcal{K} = \{1, \ldots, K\}$, is represented by a random variable $U_t^k : (\Omega, \mathscr{F}) \to (\mathcal{U}_t^k, \mathscr{U}_t^k)$, defined on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$, and takes values in the measurable space $(\mathcal{U}_t^k, \mathscr{U}_t^k)$, where $\mathcal{U}_t^k$ is subsystem k's nonempty feasible set of actions at time t and $\mathscr{U}_t^k$ is the associated σ-algebra. Let $U_t^{1:K} = (U_t^1, \ldots, U_t^K)$ be the control of CPS at time t. Starting at the initial state X_0, the evolution of the CPS model is described by the state equation

$$X_{t+1} = f_t\big(X_t, U_t^{1:K}, W_t\big), \tag{1}$$

where $t = 0, 1, \ldots, T-1$, and W_t is a random variable defined on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ that corresponds to the external, uncontrollable disturbance to the CPS and takes values in a measurable set $(\mathcal{W}, \mathscr{W})$, i.e., $W_t : (\Omega, \mathscr{F}) \to (\mathcal{W}, \mathscr{W})$. Similarly, starting at the initial state $\hat{X}_0$, the evolution of the actual CPS is described by the state equation

$$\hat{X}_{t+1} = \hat{f}_t\big(\hat{X}_t, U_t^{1:K}, W_t\big), \tag{2}$$

where $t = 0, 1, \ldots, T-1$, while $\{W_t : t = 0, \ldots, T-1\}$ is a sequence of independent random variables that are also independent of the initial states X_0 and $\hat{X}_0$. At time $t = 0, 1, \ldots, T-1$, every subsystem $k \in \mathcal{K}$ in the model makes an observation Y_t^k, which takes values in a measurable set $(\mathcal{Y}^k, \mathscr{Y}^k)$, described by the observation equation

$$Y_t^k = h_t^k(X_t, Z_t^k), \tag{3}$$

where Z_t^k is a random variable defined on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ that corresponds to the noise of each subsystem's sensor and takes values in a measurable set $(\mathcal{Z}^k, \mathscr{Z}^k)$, i.e., $Z_t^k : (\Omega, \mathscr{F}) \to (\mathcal{Z}^k, \mathscr{Z}^k)$, while $\{Z_t^k : t = 0, \ldots, T-1;\ k = 1, \ldots, K\}$ is a sequence of independent random variables that are also independent of $\{W_t : t = 0, \ldots, T-1\}$, and the initial states X_0 and $\hat{X}_0$. Similarly, at time $t = 0, 1, \ldots, T-1$, every subsystem $k \in \mathcal{K}$ in the actual CPS makes an observation $\hat{Y}_t^k$, which takes values in a measurable set $(\mathcal{Y}^k, \mathscr{Y}^k)$, described by the observation equation

$$\hat{Y}_t^k = h_t^k(\hat{X}_t, Z_t^k). \tag{4}$$

We consider that the actual CPS has n-step delayed information sharing, i.e., at time t, subsystem $k \in \mathcal{K}$ observes $\hat{Y}_t^k$, and the n-step past observations $\hat{Y}_{0:t-n}^{1:K}$ and decisions $U_{0:t-n}^{1:K}$ of the entire system. At time t, the data available to subsystem k consist of the data $\hat{\Delta}_t$ available to all subsystems, i.e.,

$$\hat{\Delta}_t := (\hat{Y}_{0:t-n}^{1:K}, U_{0:t-n}^{1:K}), \tag{5}$$

where $\hat{Y}_{0:t-n}^{1:K} = \{\hat{Y}_{0:t-n}^1, \ldots, \hat{Y}_{0:t-n}^K\}$, $U_{0:t-n}^{1:K} = \{U_{0:t-n}^1, \ldots, U_{0:t-n}^K\}$, and the data $\hat{\Lambda}_t^k$ known only to subsystem $k \in \mathcal{K}$, is given by

$$\hat{\Lambda}_t^k := (\hat{Y}_{t-n+1:t}^k, U_{t-n+1:t-1}^k). \tag{6}$$

Note that the n-step delayed information sharing can also be asymmetric, i.e., for each member $k \in \mathcal{K}$, $\hat{Y}_{t-n_k}^k$, $U_{t-n_k}^k$, where $n_k \in \mathbb{R}$ is constant but not necessarily the same for each k. The collection $\{(\hat{\Delta}_t, \hat{\Lambda}_t^k);\ k \in \mathcal{K};\ t = 0, \ldots, T-1\}$, is the information structure of the actual CPS and captures which subsystem knows what about the status of the CPS and when. In what follows, the results hold for any special case of potential information structures that can be:

(1) **Periodic information sharing with period** $\omega \geq 1$: In this case Ooi et al. (1997), for $\alpha = 1, 2, \ldots$ and $\alpha\omega < t \leq (\alpha+1)\omega$, the pair of $\hat{\Delta}_t$, $\hat{\Lambda}_t^k$, $k \in \mathcal{K}$, becomes

$$\hat{\Delta}_t := (\hat{Y}_{0:\alpha\omega}^{1:K}, U_{0:\alpha\omega}^{1:K}), \tag{7}$$

$$\hat{\Lambda}_t^k := (\hat{Y}_{\alpha\omega+1:(\alpha+1)\omega}^k, U_{\alpha\omega+1:(\alpha+1)\omega}^k). \tag{8}$$

(2) n**-step delayed observation sharing:** In this case Aicardi et al. (1987), $\hat{\Delta}_t$ and $\hat{\Lambda}_t^k$, $k \in \mathcal{K}$, become

$$\hat{\Delta}_t := (\hat{Y}_{0:t-n}^{1:K}), \tag{9}$$

$$\hat{\Lambda}_t^k := (\hat{Y}_{t-n+1:t}^k, U_{0:t-1}^k). \tag{10}$$

(3) n-**step delayed control sharing:** In this case Bismut (1973), $\hat{\Delta}_t$ and $\hat{\Lambda}_t^k$, $k \in \mathcal{K}$, become

$$\hat{\Delta}_t := (U_{0:t-n}^{1:K}), \tag{11}$$

$$\hat{\Lambda}_t^k := (\hat{Y}_{0:t}^k, U_{t-n+1:t-1}^k). \tag{12}$$

(4) **No sharing information:** In this case, $\hat{\Delta}_t$ and $\hat{\Lambda}_t^k$, $k \in \mathcal{K}$, become

$$\hat{\Delta}_t := \emptyset, \tag{13}$$

$$\hat{\Lambda}_t^k := (\hat{Y}_{0:t}^k, U_{0:t-1}^k). \tag{14}$$

The CPS model imposes the same information structure as the actual CPS. The collection $\{(\Delta_t, \Lambda_t^k);\ k \in \mathcal{K};\ t = 0, \ldots, T-1\}$, is the information structure of the model.

2.4. Optimal control problem

Let $(\mathcal{D}_t, \mathscr{D}_t)$ be the measurable spaces of all possible realizations of Δ_t and $\hat{\Delta}_t$, and $(\mathcal{L}_t^k, \mathscr{L}_t^k)$, $k \in \mathcal{K}$, be the measurable spaces of all possible realizations of Λ_t^k and $\hat{\Lambda}_t^k$, where $\mathscr{D}_t$ and $\mathscr{L}_t^k$ are the associated σ-algebras. A control strategy $\mathbf{g} = \{g_t;\ t = 0, \ldots, T-1\}$, $\mathbf{g} \in \mathcal{G}$, $\mathcal{G} =: (\mathcal{L}_t^1 \times \ldots \times \mathcal{L}_t^K \times \mathcal{D}_t, \mathscr{L}_t^1 \otimes \cdots \otimes \mathscr{L}_t^K \otimes \mathscr{D}_t)$ yields a decision

$$U_t^{1:K} = g_t(\hat{\Delta}_t, \hat{\Lambda}_t^{1:K}), \tag{15}$$

where the measurable function g_t is the control law.

Problem 1 (*Actual CPS*)**.** The problem is to derive the optimal control strategy $\mathbf{g}^* \in \mathcal{G}$ that minimizes the expected total cost of the actual CPS,

$$\hat{J}(\mathbf{g}) = \mathbb{E}^{\mathbf{g}}\left[\sum_{t=0}^{T-1} c_t(\hat{X}_t, U_t^{1:K}) + c_T(\hat{X}_T)\right], \tag{16}$$

where the expectation is with respect to the joint probability distribution of the random variables $\hat{X}_t$ and $U_t^{1:K}$ designated by the choice of $\mathbf{g} \in \mathcal{G}$, $c_t(\cdot,\cdot) : (\mathcal{X}_t \times \prod_{k\in\mathcal{K}} \mathcal{U}_t^k, \mathscr{X}_t \otimes \mathscr{U}_t^1 \otimes \cdots \otimes \mathscr{U}_t^K) \to \mathbb{R}$ is the measurable cost function of the actual CPS at t, and $c_T(\cdot) : (\mathcal{X}_T, \mathscr{X}_T) \to \mathbb{R}$ is the measurable cost function at T.

The statistics of the primitive random variables $\hat{X}_0$, $\{W_t : t = 0, \ldots, T-1\}$, $\{Z_t^k : k \in \mathcal{K};\ t = 0, \ldots, T-1\}$, the observation equations $\{h_t^k : k \in \mathcal{K};\ t = 0, \ldots, T-1\}$, and the cost functions $\{c_t : t = 0, \ldots, T\}$ are all known. However, the state equations $\{\hat{f}_t : t = 0, \ldots, T-1\}$ are not known.

3. Separation of learning and control

In our exposition, we address Problem 1 from the point of view of a central controller who seeks to derive the optimal strategy $\mathbf{g} \in \mathcal{G}$ of the actual CPS. First, we institute an appropriate information state, defined formally next, that can be used to formulate a classical dynamic programming decomposition. To establish this information state, we use the CPS model in conjunction with the actual CPS (Fig. 2).

Definition 1. An information state, Π_t, for the system described by the state Eqs. (1) and (2), is (a) a function of $(\Delta_t, \Lambda_t^{1:K})$, while (b) Π_{t+1} is determined from Π_t, $Y_{t+1}^{1:K}$, and $U_t^{1:K}$.

We consider densities for all probability distributions to simplify notation. Let $\mathbf{g} = \{g_t;\ t = 0, \ldots, T-1\}$, $\mathbf{g} \in \mathcal{G}$, be a control strategy and $(\Delta_t, \Lambda_t^{1:K})$ be the information structure of the CPS model. The control strategy $\mathbf{g}$ yields a decision $U_t^{1:K} = g_t(\Delta_t, \Lambda_t^{1:K})$.

Before we proceed with establishing the information state, we prove some essential properties.

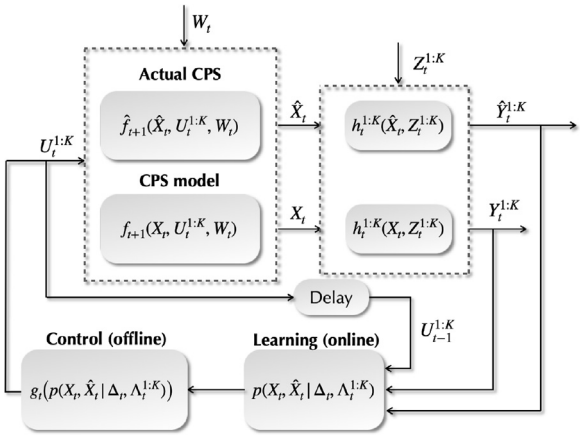

Fig. 2. Separation of learning and control.

Lemma 1. *For any control strategy* $\mathbf{g} \in \mathcal{G}$ *of the system,*

$$p^{\mathbf{g}}(Y_{t+1}^{1:K} \mid X_{t+1}, \hat{X}_{t+1}, \Delta_t, \Lambda_t^{1:K}, U_t^{1:K}) = p(Y_{t+1}^{1:K} \mid X_{t+1}), \tag{17}$$

for all $t = 0, 1, \ldots, T-1$.

Proof. The realization of $Y_{t+1}^{1:K}$ is statistically determined by the conditional distribution of $Y_{t+1}^{1:K}$ given X_{t+1} in (3), hence

$$p^{\mathbf{g}}(Y_{t+1}^{1:K} \mid X_{t+1}, \hat{X}_{t+1}, \Delta_t, \Lambda_t^{1:K}, U_t^{1:K}) = p^{\mathbf{g}}(Y_{t+1}^{1:K} \mid X_{t+1}). \tag{18}$$

However,

$$p^{\mathbf{g}}(Y_{t+1}^{1:K} \mid X_{t+1}) = p^{\mathbf{g}}(Z_{t+1}^{1:K} \in \prod_{k\in\mathcal{K}} B^k \mid X_{t+1}), \tag{19}$$

where $B^k \in \mathscr{Z}^k$, $k \in \mathcal{K}$. Since, $\{Z_t^k : k = 1, \ldots, K;\ t = 0, \ldots, T-1\}$ is a sequence of independent random variables that are independent of X_{t+1},

$$p^{\mathbf{g}}(Z_{t+1}^{1:K} \in \prod_{k\in\mathcal{K}} B^k \mid X_{t+1}) = p(Z_{t+1}^{1:K} \in \prod_{k\in\mathcal{K}} B^k). \tag{20}$$

Hence,

$$p^{\mathbf{g}}(Y_{t+1}^{1:K} \mid X_{t+1}) = p(Y_{t+1}^{1:K} \mid X_{t+1}). \tag{21}$$

The result follows from (18) and (21). □

Lemma 2. *For any control strategy* $\mathbf{g} \in \mathcal{G}$ *of the system,*

$$p^{\mathbf{g}}(X_{t+1}, \hat{X}_{t+1} \mid X_t, \hat{X}_t, \Delta_t, \Lambda_t^{1:K}, U_t^{1:K}) = p(X_{t+1}, \hat{X}_{t+1} \mid X_t, \hat{X}_t, U_t^{1:K}), \tag{22}$$

for all $t = 0, 1, \ldots, T-1$.

Proof. The realization of X_{t+1} is statistically determined by the conditional distribution of X_{t+1} given X_t and $U_t^{1:K}$, i.e., $p^{\mathbf{g}}(X_{t+1} \mid X_t, U_t^{1:K})$. Similarly, the realization of $\hat{X}_{t+1}$ is statistically determined by the conditional distribution of $\hat{X}_{t+1}$ given $\hat{X}_t$ and $U_t^{1:K}$, i.e., $p^{\mathbf{g}}(\hat{X}_{t+1} \mid \hat{X}_t, U_t^{1:K})$.

From (1), we have

$$p^{\mathbf{g}}(X_{t+1} \mid X_t, U_t^{1:K}) = p^{\mathbf{g}}(W_t \in A \mid X_t, U_t^{1:K}), \tag{23}$$

where $A \in \mathscr{W}$. From (2), we have

$$p^{\mathbf{g}}(\hat{X}_{t+1} \mid \hat{X}_t, U_t^{1:K}) = p^{\mathbf{g}}(W_t \in A \mid \hat{X}_t, U_t^{1:K}), \tag{24}$$

where $A \in \mathscr{W}$. Since, $\{W_t : t = 0, \ldots, T-1\}$ is a sequence of independent random variables that are independent of X_t, $\hat{X}_t$, and

$U_t^{1:K}$,

$$p^{\mathbf{g}}(W_t \in A \mid X_t, U_t^{1:K}) = p^{\mathbf{g}}(W_t \in A \mid \hat{X}_t, U_t^{1:K}) = p(W_t \in A). \tag{25}$$

Next,

$$p^{\mathbf{g}}(X_{t+1} \mid X_t, \Delta_t, \Lambda_t^{1:K}, U_t^{1:K}) = p^{\mathbf{g}}(W_t \in A \mid X_t, \Delta_t, \Lambda_t^{1:K}, U_t^{1:K}) = p(W_t \in A). \tag{26}$$

Similarly,

$$p^{\mathbf{g}}(\hat{X}_{t+1} \mid \hat{X}_t, \Delta_t, \Lambda_t^{1:K}, U_t^{1:K}) = p^{\mathbf{g}}(W_t \in A \mid \hat{X}_t, \Delta_t, \Lambda_t^{1:K}, U_t^{1:K}) = p(W_t \in A). \tag{27}$$

The result follows from (23), (24), (25), (26), and (27). □

Lemma 3. *For any control strategy $\mathbf{g} \in \mathcal{G}$ of the system,*

$$p^{\mathbf{g}}(X_t, \hat{X}_t \mid \Delta_t, \Lambda_t^{1:K}) = p(X_t, \hat{X}_t \mid \Delta_t, \Lambda_t^{1:K}), \tag{28}$$

for all $t = 0, 1, \ldots, T-1$.

Proof. By expanding $p^{\mathbf{g}}(X_t, \hat{X}_t \mid \Delta_t, \Lambda_t^{1:K})$, we have

$$p^{\mathbf{g}}(X_t, \hat{X}_t \mid \Delta_t, \Lambda_t^{1:K}) = p^{\mathbf{g}}(X_t, \hat{X}_t \mid \Delta_{t-1}, \Lambda_{t-2}^{1:K}, Y_{t-1}^{1:K}, Y_t^{1:K}, U_{t-2}^{1:K}, U_{t-1}^{1:K}). \tag{29}$$

However, the realizations of X_t and $\hat{X}_t$ are statistically determined by the conditional joint distribution of X_t and $\hat{X}_t$ given X_{t-1}, $\hat{X}_{t-1}$ and $U_{t-1}^{1:K}$, which does not depend on the control strategy $\mathbf{g}$ (Lemma 2), so we can drop the superscript in (29), and thus (28) follows immediately. □

Remark 1. As a consequence of Lemma 3, and since both X_t and $\hat{X}_t$ do not depend on $U_t^{1:K}$, we have

$$p^{\mathbf{g}}(X_t, \hat{X}_t \mid \Delta_t, \Lambda_t^{1:K}, U_t^{1:K}) = p(X_t, \hat{X}_t \mid \Delta_t, \Lambda_t^{1:K}). \tag{30}$$

Given that we can observe the data $(\Delta_t, \Lambda_t^{1:K})$ of the CPS model, we can compress these data to a sufficient statistic which is the probability density function $p(X_t, \hat{X}_t \mid \Delta_t, \Lambda_t^{1:K})$, called information state and denoted by $\Pi_t(\Delta_t, \Lambda_t^{1:K})(X_t, \hat{X}_t)$. The next result shows that such information state does not depend on the control strategy of the CPS model.

Theorem 1 (*Information State of the System*)**.** *For any control strategy $\mathbf{g} \in \mathcal{G}$ derived offline for the CPS model, the information state $\Pi_t(\Delta_t, \Lambda_t^{1:K})(X_t, \hat{X}_t)$ does not depend on the control strategy $\mathbf{g}$. Moreover, there is a function ϕ_t, which does not depend on the control strategy $\mathbf{g}$, such that*

$$\Pi_{t+1}(\Delta_{t+1}, \Lambda_{t+1}^{1:K})(X_{t+1}, \hat{X}_{t+1}) = \phi_t\big[\Pi_t^k(\Delta_t, \Lambda_t^{1:K})(X_t, \hat{X}_t), Y_{t+1}^{1:K}, U_t^{1:K}\big], \tag{31}$$

for all $t = 0, 1, \ldots, T-1$.

Proof. See Appendix A. □

The information state $\Pi_{t+1}(\Delta_{t+1}, \Lambda_{t+1}^{1:K})(X_{t+1}, \hat{X}_{t+1})$ of the system is the entire probability density function and not just its value at any particular realization of X_{t+1} and $\hat{X}_{t+1}$. This is because to compute $\Pi_{t+1}(\Delta_{t+1}, \Lambda_{t+1}^{1:K})(X_{t+1}, \hat{X}_{t+1})$ for any particular realization of X_{t+1} and $\hat{X}_{t+1}$, we need the probability density functions $p(\cdot, \cdot \mid \Delta_t, \Lambda_t^{1:K}, U_t^{1:K})$ and $p(\cdot, \cdot \mid \Delta_t, \Lambda_t^{1:K})$. This implies that the information state takes values in the space of these probability densities, which is an infinite-dimensional space.

In what follows, to simplify notation, the information state $\Pi_t(\Delta_t, \Lambda_t^{1:K})(X_t, \hat{X}_t)$ of the system at t is denoted simply by Π_t. We use its arguments only if it is required in our exposition.

Definition 2. A control strategy $\mathbf{g} = \{g_t;\ t = 0, \ldots, T-1\}$, of the system is said to be *separated* if g_t depends on Δ_t and $\Lambda_t^{1:K}$ only through the information state, i.e., $U_t^{1:K} = g_t\big(\Pi_t(\Delta_t, \Lambda_t^{1:K})(X_t, \hat{X}_t)\big)$. Let $\mathcal{G}^s \subseteq \mathcal{G}$ denote the set of all separated control strategies.

To derive the optimal control strategy of the actual CPS in Problem 1, we formulate the following optimization problem.

Problem 2 (*CPS Model*)**.** Using the CPS model, we seek to derive offline the optimal control strategy $\mathbf{g}^* \in \mathcal{G}^s$ that minimizes the following expected total cost

$$J(\mathbf{g}; \hat{x}_{0:T}) = \mathbb{E}^{\mathbf{g}}\Big[\sum_{t=0}^{T-1}\Big[c_t(X_t, U_t^{1:K}) + \beta \cdot |X_{t+1} - \hat{X}_{t+1}|^2\Big] + c_T(X_T)\Big], \tag{32}$$

where $X_{t+1} = f_t\big(X_t, U_t^{1:K}, W_t\big)$, $\hat{X}_{t+1} = \hat{f}_t\ \big(\ \hat{X}_t, U_t^{1:K}, W_t\ \big)$, and β is a factor to adjust the units and size of the norm accordingly as designated by the cost function $c_t(\cdot, \cdot)$. The norm penalizes any discrepancy between the realizations of the state of the CPS model and the state of the actual CPS. The expectation in (32) is with respect to the joint probability distribution of the random variables X_t, $U_t^{1:K}$, $\hat{X}_t$, $t = 0, 1, \ldots, T$, (designated by the choice of $\mathbf{g} \in \mathcal{G}^s$) and W_t. Since solving (32) is an offline process, the realizations $\hat{x}_{0:T}$ of the state $\hat{X}_t$, $t = 0, \ldots, T$, of the actual CPS are not known, and thus $\mathbf{g}^*$ is parameterized with respect to $\hat{x}_{0:T}$. The statistics of the primitive random variables X_0, $\{W_t : t = 0, \ldots, T-1\}$, $\{Z_t^k : k \in \mathcal{K};\ t = 0, \ldots, T-1\}$, the state equations $\{f_t : t = 0, \ldots, T-1\}$, the observation equations $\{h_t^k : k \in \mathcal{K};\ t = 0, \ldots, T-1\}$, and the cost functions $\{c_t : t = 0, \ldots, T\}$ are all known.

Next, we use the information state $\Pi_t(\Delta_t, \Lambda_t^{1:K})\ (X_t, \hat{X}_t)$ to derive offline the optimal separated control strategy in Problem 2. In our exposition, we define recursive functions, and show that a separated control strategy of the CPS model is optimal. In addition, we obtain a classical dynamic programming decomposition.

Theorem 2. *Let $V_t\big(\Pi_t(\Delta_t, \Lambda_t^{1:K})(X_t, \hat{X}_t)\big)$ be functions defined recursively for all $\mathbf{g} \in \mathcal{G}^s$ by*

$$V_T\big(\Pi_T(\Delta_T, \Lambda_T^{1:K})(X_T, \hat{X}_T)\big) := \mathbb{E}^{\mathbf{g}}\Big[c_T(X_T) \mid \Pi_T = \pi_T\Big],$$

$$V_t\big(\Pi_t(\Delta_t, \Lambda_t^{1:K})(X_t, \hat{X}_t)\big) := \inf_{u_t^{1:K} \in \prod_{k \in \mathcal{K}} \mathcal{U}_t^k} \mathbb{E}^{\mathbf{g}}\Big[c_t(X_t, U_t^{1:K}) + \beta\ |X_{t+1} - \hat{X}_{t+1}|^2 + V_{t+1}\big(\phi_t\big[\Pi_t(\Delta_t, \Lambda_t^{1:K})(X_t, \hat{X}_t), Y_{t+1}^{1:K}, U_t^{1:K}\big]\big) \mid \Pi_t = \pi_t, U_t^{1:K} = u_t^{1:K}\Big], \tag{33}$$

where $c_T(X_T)$ is the cost function at T; β is a factor to adjust the units and size of the norm as designated by the cost function $c_t(\cdot, \cdot)$; and π_T, π_t, $u_t^{1:K}$ are the realizations of Π_T, Π_t, and $U_t^{1:K}$, respectively. Then, (a) for any control strategy $\mathbf{g} \in \mathcal{G}^s$,

$$V_t\big(\Pi_t(\Delta_t, \Lambda_t^{1:K})(X_t, \hat{X}_t)\big) \le J_t(\mathbf{g}; \hat{x}_{t:T}) := \mathbb{E}^{\mathbf{g}}\Big[\sum_{l=t}^{T-1}\Big[c_l(X_l, U_l^{1:K}) + \beta \cdot |X_{l+1} - \hat{X}_{l+1}|^2\Big] + c_T(X_T) \mid \Delta_t, \Lambda_t^{1:K}\Big], \tag{34}$$

where $J_t(\mathbf{g}; \hat{x}_{t:T})$ is the cost-to-go function of the CPS model, parameterized by the realizations of the state $\hat{X}_t$ of the actual CPS, at time t corresponding to the control strategy $\mathbf{g}$; and (b) $\mathbf{g} \in \mathcal{G}^s$ is optimal and

$$V_t\big(\Pi_t(\Delta_t, \mathcal{I}_t^{1:K})(X_t, \hat{X}_t)\big) = J_t(\mathbf{g}; \hat{x}_{t:T}), \tag{35}$$

with probability 1.

Proof. See Appendix B. □

The optimal strategy derived by the CPS model, which is parameterized with respect to the potential realizations $\hat{x}_{0:T}$ of the state $\hat{X}_t$, $t = 0, \ldots, T$, of the actual CPS, is used to operate the actual CPS in parallel with the CPS model (Fig. 2). As we collect data from the two systems, we learn the information state $\Pi_t(\Delta_t, \mathcal{I}_t^{1:K})(X_t, \hat{X}_t)$ online.

Proposition 1. *The information state $\Pi_t(\Delta_t, \mathcal{I}_t^{1:K})(X_t, \hat{X}_t)$ of the system is a function of $p(X_t \mid \Delta_t, \mathcal{I}_t^{1:K})$, $p(\hat{X}_t \mid \hat{\Delta}_t, \hat{\mathcal{I}}_t^{1:K})$, and $p(\hat{Y}_{0:t}^{1:K} \mid U_{0:t-1}^{1:K})$.*

Proof. Recall $\Pi_t(\Delta_t, \mathcal{I}_t^{1:K})(X_t, \hat{X}_t) = p(X_t, \hat{X}_t \mid \Delta_t, \mathcal{I}_t^{1:K})$. Next,

$$\begin{aligned}
&p(X_t, \hat{X}_t \mid \Delta_t, \mathcal{I}_t^{1:K}) \\
&= \frac{p(\hat{X}_t \mid X_t, \Delta_t, \mathcal{I}_t^{1:K}) \cdot p(X_t, \Delta_t, \mathcal{I}_t^{1:K})}{p(\Delta_t, \mathcal{I}_t^{1:K})} \\
&= \frac{p(\hat{X}_t \mid U_{0:t-1}^{1:K}) \cdot p(X_t, \Delta_t, \mathcal{I}_t^{1:K})}{p(\Delta_t, \mathcal{I}_t^{1:K})} \\
&= p(\hat{X}_t \mid U_{0:t-1}^{1:K}) \cdot p(X_t \mid \Delta_t, \mathcal{I}_t^{1:K}),
\end{aligned} \tag{36}$$

where, in the second equality, we used the fact that $\hat{X}_t$ does not depend on X_t and $Y_{0:t}^{1:K}$, and in the third equality we applied Bayes' rule. The first term in (36) can be written as

$$\begin{aligned}
&p(\hat{X}_t \mid U_{0:t-1}^{1:K}) \\
&= \int_{\mathscr{X}_t} p(\hat{X}_t \mid \hat{Y}_{0:t}^{1:K}, U_{0:t-1}^{1:K}) \cdot p(\hat{Y}_{0:t}^{1:K} \mid U_{0:t-1}^{1:K}) d\hat{Y}_{0:t}^{1:K},
\end{aligned} \tag{37}$$

and the result follows. □

Remark 2. The conditional probabilities $p(X_t \mid \Delta_t, \mathcal{I}_t^{1:K})$ and $p(\hat{X}_t \mid \hat{\Delta}_t, \hat{\mathcal{I}}_t^{1:K})$ can be computed recursively starting from an initial prior $p(X_0 \mid \Delta_0, \mathcal{I}_0^{1:K})$ and $p(\hat{X}_0 \mid \hat{\Delta}_0, \hat{\mathcal{I}}_0^{1:K})$,

$$\begin{aligned}
&p(X_t \mid \Delta_t, \mathcal{I}_t^{1:K}) \\
&= \theta_{t-1}\big[p(X_{t-1} \mid \Delta_{t-1}, \mathcal{I}_{t-1}^{1:K}), Y_t^{1:K}, U_{t-1}^{1:K}\big],
\end{aligned} \tag{38}$$

$$\begin{aligned}
&p(\hat{X}_t \mid \hat{\Delta}_t, \hat{\mathcal{I}}_t^{1:K}) \\
&= \hat{\theta}_{t-1}\big[p(\hat{X}_{t-1} \mid \hat{\Delta}_{t-1}, \hat{\mathcal{I}}_{t-1}^{1:K}), Y_t^{1:K}, U_{t-1}^{1:K}\big],
\end{aligned} \tag{39}$$

for all $t = 0, 1, \ldots, T-1$, where θ_t and $\hat{\theta}_t$ are appropriate functions; see Malikopoulos (2023).

Remark 3. The information state $\Pi_t(\Delta_t, \mathcal{I}_t^{1:K})(X_t, \hat{X}_t)$ of the system can be obtained by using standard learning approaches, i.e., Brand (1999), Gyorfi and Kohler (2007), to learn online the conditional probabilities $p(\hat{Y}_{0:t}^{1:K} \mid U_{0:t-1}^{1:K})$ while we operate the actual CPS.

Next, we show that after the information state becomes known through learning, then the separated control strategy of the CPS model derived offline is optimal for the actual CPS.

Theorem 3. *Let $\mathbf{g} \in \mathcal{G}^s$ be an optimal separated control strategy derived offline for the CPS model which minimizes the expected total cost,*

$$\begin{aligned}
J(\mathbf{g}; \hat{x}_{0:T}) := \mathbb{E}^{\mathbf{g}}\Bigg[\sum_{t=0}^{T-1}\Big[c_t(X_t, U_t^{1:K}) + \beta \cdot |X_{t+1} \\
- \hat{X}_{t+1}|^2\Big] + c_T(X_T)\Bigg],
\end{aligned} \tag{40}$$

in Problem 2. If $p(X_t, \hat{X}_t \mid \Delta_t, \mathcal{I}_t^{1:K}) = \Pi(\Delta_t, \mathcal{I}_t^{1:K})(X_{t+1}, \hat{X}_{t+1})$ is known, then $\mathbf{g}$ minimizes also the expected total cost of the actual CPS,

$$\hat{J}(\mathbf{g}) = \mathbb{E}^{\mathbf{g}}\left[\sum_{t=0}^{T-1} c_t(\hat{X}_t, U_t^{1:K}) + c_T(\hat{X}_T)\right], \tag{41}$$

in Problem 1.

Proof. If $p(X_t, \hat{X}_t \mid \Delta_t, \mathcal{I}_t^{1:K}) = \Pi(\Delta_t, \mathcal{I}_t^{1:K})(X_{t+1}, \hat{X}_{t+1})$ is known, then, for all $t = 0, \ldots, T-1$, $U_t^{1:K} = g_t\big(\Pi(\Delta_t, \mathcal{I}_t^{1:K})(X_{t+1}, \hat{X}_{t+1})\big)$ minimizes (40), which implies

$$|X_{t+1} - \hat{X}_{t+1}|^2 = 0, \tag{42}$$

for all $t = 0, \ldots, T-1$, hence $c_t(X_t, U_l^{1:K}) = c_t(\hat{X}_t, U_l^{1:K})$ and $c_T(X_T) = c_T(\hat{X}_T)$. Therefore,

$$\begin{aligned}
J(\mathbf{g}; \hat{x}_{0:T}) &= \mathbb{E}^{\mathbf{g}}\left[\sum_{t=0}^{T-1} c_t(X_t, U_t^{1:K}) + c_T(X_T)\right] \\
&= \mathbb{E}^{\mathbf{g}}\left[\sum_{t=0}^{T-1} c_t(\hat{X}_t, U_t^{1:K}) + c_T(\hat{X}_T)\right] = \hat{J}(\mathbf{g}).
\end{aligned} \tag{43}$$

□

The following results provide some structural properties of the recursive functions.

Lemma 4. *The function $V_t\big(\Pi_t(\Delta_t, \mathcal{I}_t^{1:K})(X_{t+1}, \hat{X}_{t+1})\big)$ defined recursively in Theorem 2 is positive homogeneous for all $t = 0, \ldots, T$, i.e., for any $\rho > 0$, $V_t\big(\rho\ \Pi_t(\Delta_t, \mathcal{I}_t^{1:K})(X_{t+1}, \hat{X}_{t+1})\big) = \rho\ V_t\ \big(\Pi_t(\Delta_t, \mathcal{I}_t^{1:K})(X_{t+1}, \hat{X}_{t+1})\big)$.*

Proof. See Appendix C. □

Theorem 4. *The function $V_t\big(\Pi_t(\Delta_t, \mathcal{I}_t^{1:K})(X_{t+1}, \hat{X}_{t+1})\big)$ defined recursively in Theorem 2 is concave with respect to $\Pi_t(\Delta_t, \mathcal{I}_t^{1:K})(X_{t+1}, \hat{X}_{t+1})$.*

Proof. See Appendix D. □

Remark 4. From Theorem 4, the solution of Problem 2 can be derived using standard techniques for centralized partially observed Markov decision processes. If the observation space of the CPS is finite, then (32) has a finite dimensional characterization (see Krishnamurthy (2016), p. 154). In particular, the explicit solution to (32) is a piecewise linear concave function of the information state; see Sondik (1971).

4. Illustrative example

We present a simple example of a system consisting of two subsystems ($K = 2$) with delayed sharing pattern to illustrate the proposed framework. The system evolves for a time horizon $T = 4$ while there is a delay $n = 2$ on information sharing between the two subsystems. The state of the actual system $\hat{X}_t = (\hat{X}_t^1, \hat{X}_t^2)$, $t = 1, 2, 3, 4$, is two-dimensional, and the initial state (primitive random variable), $\hat{X}_0 = (\hat{X}_0^1, \hat{X}_0^2)$, of the system is a Gaussian random variable with zero mean, variance 1, and covariance 0.5.

The state of the actual system evolves as follows:

$$\hat{X}_0 = (\hat{X}_0^1, \hat{X}_0^2), \tag{44}$$
$$\hat{X}_1 = (\hat{X}_1^1, \hat{X}_1^2) = (\hat{X}_0^1, \hat{X}_0^2), \tag{45}$$
$$\hat{X}_2 = (\hat{X}_2^1, \hat{X}_2^2) = (\hat{X}_0^1 + \hat{X}_0^2, 0), \tag{46}$$
$$\hat{X}_3 = (\hat{X}_3^1, \hat{X}_3^2) = (\hat{X}_2^1, U_2^2) = (\hat{X}_0^1 + \hat{X}_0^2, U_2^2), \tag{47}$$
$$\hat{X}_4 = (\hat{X}_4^1, \hat{X}_4^2) = (\hat{X}_3^1 - \hat{X}_3^2 - U_3^1, 0) = (\hat{X}_0^1 + \hat{X}_0^2 - U_2^2 - U_3^1, 0), \tag{48}$$

and the observation equations are

$$\hat{Y}_t^k = \hat{X}_t^k, \quad k = 1, 2; \ t = 1, 2, 3, 4. \tag{49}$$

The state of the system's model evolves as follows:

$$X_0 = (X_0^1, X_0^2), \tag{50}$$
$$X_1 = (X_1^1, X_1^2) = (X_0^1, X_0^2), \tag{51}$$
$$X_2 = (X_2^1, X_2^2) = (X_0^1 + X_0^2, 0), \tag{52}$$
$$X_3 = (X_3^1, X_3^2) = (X_2^1, U_2^2) = (X_0^1 + X_0^2, U_2^2), \tag{53}$$
$$X_4 = (X_4^1, X_4^2) = (X_3^1 - X_3^2 - U_3^1, 0) = (X_0^1 + X_0^2 - U_2^2 - U_3^1, 0), \tag{54}$$

and the observation equations are

$$Y_t^k = X_t^k, \quad k = 1, 2; \ t = 1, 2, 3, 4. \tag{55}$$

Since X_0 and $\hat{X}_0$ are different, we have implicitly imposed an artificial discrepancy between the model and the actual system.

Each subsystem's feasible sets of actions $\mathcal{U}_t^k$ are specified by

$$\mathcal{U}_t^k = \begin{cases} \mathbb{R}, & \text{if } (k,t) = (1,3) \text{ or } (2,2), \\ 0, & \text{otherwise.} \end{cases} \tag{56}$$

Hence a control strategy $\mathbf{g} \in \mathcal{G}^s$ of the system consists only of the pair $\mathbf{g} = \{g_2, g_3\}$ since $g_t \equiv 0$ for the remaining t. Given the modeling framework above, the information structure $\{(\Delta_t, \Lambda_t^k); k = 1, 2; \ t = 1, 2, 3\}$ of the system is captured through the model as follows

$$\Delta_1 = \emptyset, \ \Delta_2 = \emptyset, \tag{57}$$
$$\Delta_3 = \{Y_0^1, Y_0^2, Y_1^1, Y_1^2\} = \{X_0^1, X_0^2\}. \tag{58}$$

Note that since $g_1 \equiv 0$, the realizations of U_1^1 and U_1^2 are zero, and thus Δ_3 includes only the observations in (58). The data Λ_t^k, $k = 1, 2$, available to subsystem k for the feasible control laws are

$$\Lambda_2^1 = \{Y_0^1, Y_1^1, Y_2^1\} = \{X_0^1, X_1^1, X_2^1\} = \{X_0^1, X_0^1 + X_0^2\}, \tag{59}$$
$$\Lambda_2^2 = \{Y_0^2, Y_1^2, Y_2^2\} = \{X_0^2, X_1^2, X_2^2\} = \{X_0^2\}, \tag{60}$$
$$\Lambda_3^1 = \{Y_2^1, Y_3^1\} = \{X_0^1 + X_0^2, X_0^1 + X_0^2\} = \{X_0^1 + X_0^2\}, \tag{61}$$
$$\Lambda_3^2 = \{Y_2^2, Y_3^2, U_2^2\} = \{U_2^2\}. \tag{62}$$

4.1. Optimal solution

The problem is to derive the optimal control strategy $\mathbf{g}^* \in \mathcal{G}^s$ of the actual system which is the solution of

$$J(\mathbf{g}) = \min_{u_2^2 \in \mathcal{U}_2^2, u_3^1 \in \mathcal{U}_3^1} \frac{1}{2}\mathbb{E}^{\mathbf{g}}\left[(\hat{X}_3^1)^2 + (U_3^1)^2\right] = \min_{u_2^2 \in \mathcal{U}_2^2, u_3^1 \in \mathcal{U}_3^1} \frac{1}{2}\mathbb{E}^{\mathbf{g}}\left[(\hat{X}_0^1 + \hat{X}_0^2 - U_2^2 - U_3^1)^2 + (U_3^1)^2\right]. \tag{63}$$

The feasible set $\mathcal{G}$ of the control strategies of the system consists of all $\mathbf{g} = \{g_2(\Lambda_2^2, \Delta_2), g_3(\Lambda_3^1, \Delta_3)\}$, i.e.,

$$g_2^2 : \Delta_2 \times \Lambda_2^2 \to U_2^2, \quad \text{or} \quad g_2^2 : \hat{X}_0^2 \to \mathbb{R}, \tag{64}$$
$$g_3^1 : \Delta_3 \times \Lambda_3^1 \to U_3^1, \quad \text{or} \quad g_3^1 : \{\hat{X}_0^1, \hat{X}_0^2\} \to \mathbb{R}. \tag{65}$$

The problem (63) has a unique optimal solution

$$U_2^2 = \frac{1}{2}\hat{X}_0^2, \quad U_3^1 = \frac{1}{2}(\hat{X}_0^1 + \hat{X}_0^2) - \frac{1}{4}\hat{X}_0^2. \tag{66}$$

4.2. Solution given by Theorems 2 and 3

We solve problem (63) by considering the control strategies $\mathbf{g} = \{g_t; \ t = 0, 1, 2, 3\}$, where the control law is of the form $g_t\left(\Pi(\Delta_t, \Lambda_t^{1:2})(X_t, \hat{X}_t)\right) = g_t\left(\mathbb{P}(X_t, \hat{X}_t \mid \Delta_t, \Lambda_t^{1:2})\right)$.

For $t = 3$, using (33) with $\beta = 1$, we have

$$V_3(\Pi_3) = \min_{u_2^2 \in \mathcal{U}_2^2, u_3^1 \in \mathcal{U}_3^1} \frac{1}{2}\mathbb{E}^{\mathbf{g}}\Big[(X_0^1 + X_0^2 - U_2^2 - U_3^1)^2 + (U_3^1)^2 + |X_4 - \hat{X}_4|^2 \mid \Pi_3\big(\Delta_3, \Lambda_3^{1:2}\big), U_3^{1:2}\Big]$$
$$= \min_{u_2^2 \in \mathcal{U}_2^2, u_3^1 \in \mathcal{U}_3^1} \frac{1}{2}\mathbb{E}^{\mathbf{g}}\Big[(X_0^1 + X_0^2 - U_2^2 - U_3^1)^2 + (U_3^1)^2 + |X_0^1 + X_0^2 - \hat{X}_0^1 - \hat{X}_0^2|^2 \mid \mathbb{P}(X_0^1 + X_0^2, U_2^2, \hat{X}_0^1 + \hat{X}_0^2 \mid X_0^1, X_0^2, X_0^1 + X_0^2, U_2^2), U_3^1\Big], \tag{67}$$

where, given the information state Π_3, we can select the realization of U_3^1 that achieves the lower bound in (67). Hence,

$$U_3^1 = \frac{1}{2}(X_0^1 + X_0^2) - \frac{1}{2}U_2^2. \tag{68}$$

Substituting (68) into (67) yields

$$V_3(\Pi_3) = \min_{u_2^2 \in \mathcal{U}_2^2, u_3^1 \in \mathcal{U}_3^1} \frac{1}{2}\mathbb{E}^{\mathbf{g}}\Big[\frac{\left(X_0^1 + X_0^2 - U_2^2\right)^2}{2} + |X_0^1 + X_0^2 - \hat{X}_0^1 - \hat{X}_0^2|^2 \mid \mathbb{P}(X_0^1 + X_0^2, U_2^2, \hat{X}_0^1 + \hat{X}_0^2 \mid X_0^1, X_0^2, X_0^1 + X_0^2, U_2^2), U_3^1\Big]. \tag{69}$$

For $t = 2$, using (33) with $\beta = 1$, we have

$$V_2(\Pi_2) = \min_{u_2^2 \in \mathcal{U}_2^2, u_3^1 \in \mathcal{U}_3^1} \frac{1}{2}\mathbb{E}^{\mathbf{g}}\Big[V_3(\Pi_3) + |X_3 - \hat{X}_3|^2 \mid \Pi_2\big(\Delta_2, \Lambda_2^{1:2}\big), U_2^{1:2}\Big]$$
$$= \min_{u_2^2 \in \mathcal{U}_2^2, u_3^1 \in \mathcal{U}_3^1} \frac{1}{2}\mathbb{E}^{\mathbf{g}}\Big[V_3(\Pi_3) + |X_0^1 + X_0^2 - \hat{X}_0^1 - \hat{X}_0^2|^2 \mid \mathbb{P}(X_0^1 + X_0^2, \hat{X}_0^1 + \hat{X}_0^2 \mid X_0^1, X_0^1 + X_0^2, X_0^2), U_2^2\Big] \tag{70}$$
$$= \min_{u_2^2 \in \mathcal{U}_2^2, u_3^1 \in \mathcal{U}_3^1} \frac{1}{2}\mathbb{E}^{\mathbf{g}}\Big[\frac{\left(X_0^1 + X_0^2 - U_2^2\right)^2}{2} + 2 \cdot |X_0^1 + X_0^2 - \hat{X}_0^1 - \hat{X}_0^2|^2 \mid \mathbb{P}(X_0^1 + X_0^2, \hat{X}_0^1 + \hat{X}_0^2 \mid X_0^1, X_0^1 + X_0^2, X_0^2), U_2^2\Big]. \tag{71}$$

Since

$$U_2^2 = g_2\big(\mathbb{P}(X_2, \hat{X}_2 \mid \Delta_2, \Lambda_2^{1:2})\big) = g_2^2\big(\mathbb{P}(X_0^1 + X_0^2, \hat{X}_0^1 + \hat{X}_0^2 \mid X_0^1, X_0^1 + X_0^2, X_0^2)\big), \tag{72}$$

the problem is to choose, for any given X_0^2, the estimate of $(X_0^1 + X_0^2)$ that minimizes the mean squared error $\left(X_0^1 + X_0^2 - U_2^2\right)^2$ in (71).

Given the Gaussian statistics, the optimal solution is

$$U_2^2 = \frac{1}{2}X_0^2. \tag{73}$$

Substituting (73) into (68) yields

$$U_3^1 = \frac{1}{2}(X_0^1 + X_0^2) - \frac{1}{4}X_0^2. \tag{74}$$

After learning the information states $\Pi_3(\Delta_3, \Lambda_3^{1:2})$ and $\Pi_2(\Delta_2, \Lambda_2^{1:2})$, the "true" values of the initial states in (73) and (74) corresponding to the actual system become known. Hence, we select $X_0^1 = \hat{X}_0^1$ and $X_0^2 = \hat{X}_0^2$, and thus $U_2^2 = \frac{1}{2}\hat{X}_0^2$ and $U_3^1 = \frac{1}{2}(\hat{X}_0^1 + \hat{X}_0^2) - \frac{1}{4}\hat{X}_0^2$. Therefore, the control laws of the form $g_t\big(\Pi(\Delta_t, \Lambda_t^{1:K})(X_t, \hat{X}_t)\big) = g_t\big(\mathbb{P}(X_t, \hat{X}_t \mid \Delta_t, \Lambda_t^{1:K})\big)$ yield the unique optimal solution (66) of problem (63).

5. Concluding remarks and discussion

In most CPS applications there is a large volume of data of a dynamic nature which is added to the system gradually in real time and not altogether in advance. As the volume of data increases, the domain of the control strategies also increases, and thus it becomes challenging to search for an optimal strategy. Even if an optimal strategy is found, implementing such strategies with increasing domains is burdensome. In such CPS applications, we typically assume an ideal model of the system which is used to derive the optimal control strategy. Such model-based control approaches cannot effectively facilitate optimal solutions with performance guarantees due to the discrepancy between the model and the actual CPS. On the other hand, traditional supervised learning approaches cannot always facilitate robust solutions using data derived offline. By contrast, applying reinforcement learning approaches directly to the actual CPS might impose significant implications on safety and robust operation of the system.

In this paper, we presented a theoretical framework that circumvents these challenges. The framework can combine offline model-based control with online learning approaches to yield the optimal control strategy for the system. There are two features which sharply distinguish the framework presented here from previous learning-based, or combined learning and control approaches reported in the literature to date: (1) the CPS imposes a nonclassical information structure while the state of the system is not fully observed; and (2) the large volume of data that is added to the system gradually is compressed to a sufficient information state without loss of optimality that takes values in a time-invariant space. Therefore, the volume of data which is added to the system gradually does not lead the domain of the control strategies to increase with time.

In our exposition, we restricted attention to centralized strategies. Ongoing research includes expanding the framework to decentralized strategies. A direction of future research should consider investigating how potential errors in the communication between the subsystems could be addressed.

Appendix A. Proof of Theorem 1

By applying Bayes' rule, we have

$$p^{\mathbf{g}}(X_{t+1}, \hat{X}_{t+1} \mid \Delta_{t+1}, \Lambda_{t+1}^{1:K}) = \frac{p^{\mathbf{g}}(Y_{t+1}^{1:K} \mid X_{t+1}, \hat{X}_{t+1}, \Delta_{t+1}, \Lambda_t^{1:K}, U_t^{1:K}) \cdot p^{\mathbf{g}}(X_{t+1}, \hat{X}_{t+1}, \Delta_{t+1}, \Lambda_t^{1:K}, U_t^{1:K})}{p^{\mathbf{g}}(\Delta_{t+1}, \Lambda_{t+1}^{1:K})} \tag{A.1}$$

$$= \frac{p(Y_{t+1}^{1:K} \mid X_{t+1})\, p^{\mathbf{g}}(X_{t+1}, \hat{X}_{t+1}, \Delta_{t+1}, \Lambda_t^{1:K}, U_t^{1:K})}{p^{\mathbf{g}}(\Delta_{t+1}, \Lambda_{t+1}^{1:K})}$$

$$= \frac{p(Y_{t+1}^{1:K} \mid X_{t+1})\, p^{\mathbf{g}}(X_{t+1}, \hat{X}_{t+1} \mid \Delta_{t+1}, \Lambda_t^{1:K}, U_t^{1:K}) \cdot p^{\mathbf{g}}(\Delta_{t+1}, \Lambda_t^{1:K}, U_t^{1:K})}{p^{\mathbf{g}}(\Delta_{t+1}, \Lambda_{t+1}^{1:K})}, \tag{A.2}$$

where in the second equality we used Lemma 1.

Next,

$$\begin{aligned}
p^{\mathbf{g}}(\Delta_{t+1}, \Lambda_{t+1}^{1:K}) &= p^{\mathbf{g}}(\Delta_{t+1}, \Lambda_t^{1:K}, Y_{t+1}^{1:K}, U_t^{1:K}) \\
&= \int_{\mathscr{X}_{t+1}} \int_{\mathscr{X}_{t+1}} p^{\mathbf{g}}(X_{t+1}, \hat{X}_{t+1}, \Delta_{t+1}, \Lambda_t^{1:K}, Y_{t+1}^{1:K}, U_t^{1:K})\, dX_{t+1}\, d\hat{X}_{t+1} \\
&= \int_{\mathscr{X}_{t+1}} \int_{\mathscr{X}_{t+1}} p^{\mathbf{g}}(Y_{t+1}^{1:K} \mid X_{t+1}, \hat{X}_{t+1}, \Delta_{t+1}, \Lambda_t^{1:K}, U_t^{1:K}) \cdot p^{\mathbf{g}}(X_{t+1}, \hat{X}_{t+1}, \Delta_{t+1}, \Lambda_t^{1:K}, U_t^{1:K})\, dX_{t+1}\, d\hat{X}_{t+1} \\
&= \int_{\mathscr{X}_{t+1}} \int_{\mathscr{X}_{t+1}} p^{\mathbf{g}}(Y_{t+1}^{1:K} \mid X_{t+1}, \hat{X}_{t+1}, \Delta_{t+1}, \Lambda_t^{1:K}, U_t^{1:K}) \cdot p^{\mathbf{g}}(X_{t+1}, \hat{X}_{t+1} \mid \Delta_{t+1}, \Lambda_t^{1:K}, U_t^{1:K}) \\
&\quad \cdot p^{\mathbf{g}}(\Delta_{t+1}, \Lambda_t^{1:K}, U_t^{1:K})\, dX_{t+1}\, d\hat{X}_{t+1},
\end{aligned}$$

where by Lemma 1, the last equation becomes

$$\begin{aligned}
p^{\mathbf{g}}(\Delta_{t+1}, \Lambda_{t+1}^{1:K}) &= \int_{\mathscr{X}_{t+1}} \int_{\mathscr{X}_{t+1}} p(Y_{t+1}^{1:K} \mid X_{t+1})\, p^{\mathbf{g}}(X_{t+1}, \hat{X}_{t+1} \mid \Delta_{t+1}, \\
&\quad \Lambda_t^{1:K}, U_t^{1:K}) \cdot p^{\mathbf{g}}(\Delta_{t+1}, \Lambda_t^{1:K}, U_t^{1:K})\, dX_{t+1}\, d\hat{X}_{t+1}.
\end{aligned} \tag{A.3}$$

Note that $p^{\mathbf{g}}(X_{t+1}, \hat{X}_{t+1} \mid \Delta_{t+1}, \Lambda_t^{1:K}, U_t^{1:K}) = p^{\mathbf{g}}(X_{t+1}, \hat{X}_{t+1} \mid \Delta_t, \Lambda_t^{1:K}, U_t^{1:K})$ since $Y_{t-n+1}^{1:K}$ and $U_{t-n+1}^{1:K}$ are already included in $\Lambda_t^{1:K}$, hence we can write (A.3) as

$$\begin{aligned}
p^{\mathbf{g}}(\Delta_{t+1}, \Lambda_{t+1}^{1:K}) &= \int_{\mathscr{X}_{t+1}} \int_{\mathscr{X}_{t+1}} p(Y_{t+1}^{1:K} \mid X_{t+1})\, p^{\mathbf{g}}(X_{t+1}, \hat{X}_{t+1} \mid \Delta_t, \\
&\quad \Lambda_t^{1:K}, U_t^{1:K}) \cdot p^{\mathbf{g}}(\Delta_{t+1}, \Lambda_t^{1:K}, U_t^{1:K})\, dX_{t+1}\, d\hat{X}_{t+1}.
\end{aligned} \tag{A.4}$$

Substituting (A.4) into (A.2), we have

$$p^{\mathbf{g}}(X_{t+1}, \hat{X}_{t+1} \mid \Delta_{t+1}, \Lambda_{t+1}^{1:K}) = \frac{p(Y_{t+1}^{1:K} \mid X_{t+1})\, p^{\mathbf{g}}(X_{t+1}, \hat{X}_{t+1} \mid \Delta_t, \Lambda_t^{1:K}, U_t^{1:K})}{\int_{\mathscr{X}_{t+1}} \int_{\mathscr{X}_{t+1}} p(Y_{t+1}^{1:K} \mid X_{t+1})\, p^{\mathbf{g}}(X_{t+1}, \hat{X}_{t+1} \mid \Delta_t, \Lambda_t^{1:K}, U_t^{1:K})\, dX_{t+1}\, d\hat{X}_{t+1}}. \tag{A.5}$$

Next,

$$\begin{aligned}
p^{\mathbf{g}}(X_{t+1}, \hat{X}_{t+1} \mid \Delta_t, \Lambda_t^{1:K}, U_t^{1:K}) &= \int_{\mathscr{X}_t} \int_{\mathscr{X}_t} p^{\mathbf{g}}(X_{t+1}, \hat{X}_{t+1} \mid X_t, \hat{X}_t, \Delta_t, \Lambda_t^{1:K}, U_t^{1:K}) \\
&\quad \cdot p^{\mathbf{g}}(X_t, \hat{X}_t \mid \Delta_t, \Lambda_t^{1:K}, U_t^{1:K})\, dX_t\, d\hat{X}_t.
\end{aligned} \tag{A.6}$$

By Lemma 2 and Remark 1, (A.6) becomes

$$\begin{aligned}
p^{\mathbf{g}}(X_{t+1}, \hat{X}_{t+1} \mid \Delta_t, \Lambda_t^{1:K}, U_t^{1:K}) &= \int_{\mathscr{X}_t} \int_{\mathscr{X}_t} p(X_{t+1}, \hat{X}_{t+1} \mid X_t, \hat{X}_t, U_t^{1:K}) \\
&\quad \cdot p(X_t, \hat{X}_t \mid \Delta_t, \Lambda_t^{1:K})\, dX_t\, d\hat{X}_t.
\end{aligned} \tag{A.7}$$

Substituting (A.7) into (A.5) yields

$$p^{\mathbf{g}}(X_{t+1}, \hat{X}_{t+1} \mid \Delta_{t+1}, \Lambda_{t+1}^{1:K}) = \frac{p(Y_{t+1}^{1:K} \mid X_{t+1}) \int_{\mathscr{X}_t} \int_{\mathscr{X}_t} p(X_{t+1}, \hat{X}_{t+1} \mid X_t, \hat{X}_t, U_t^{1:K}) \cdot p(X_t, \hat{X}_t \mid \Delta_t, \Lambda_t^{1:K})\, dX_t\, d\hat{X}_t}{\int_{\mathscr{X}_{t+1}} \int_{\mathscr{X}_{t+1}} p(Y_{t+1}^{1:K} \mid X_{t+1}) \int_{\mathscr{X}_t} \int_{\mathscr{X}_t} p(X_{t+1}, \hat{X}_{t+1} \mid X_t, U_t^{1:K})\, p(X_t, \hat{X}_t \mid \Delta_t, \Lambda_t^{1:K})\, dX_t\, d\hat{X}_t\, dX_{t+1}\, d\hat{X}_{t+1}}. \tag{A.8}$$

Therefore, $p^{\mathbf{g}}(X_{t+1}, \hat{X}_{t+1} \mid \Delta_{t+1}, \Lambda_{t+1}^{1:K})$ does not depend on the control strategy $\mathbf{g}$, so we can drop the superscript. Moreover, we can choose appropriate function ϕ_t such that

$$\begin{aligned} &p^{\mathbf{g}}(X_{t+1}, \hat{X}_{t+1} \mid \Delta_{t+1}, \Lambda_{t+1}^{1:K}) \\ &= \Pi_{t+1}(\Delta_{t+1}, \Lambda_{t+1}^{1:K})(X_{t+1}, \hat{X}_{t+1}) \\ &= \phi_t \Big[\Pi_t(\Delta_t, \Lambda_t^{1:K})(X_t, \hat{X}_t), Y_{t+1}^{1:K}, U_t^{1:K} \Big]. \end{aligned} \tag{A.9}$$

Appendix B. Proof of Theorem 2

(a) We prove (34) by induction. For $t = T$,

$$\begin{aligned} &J_T(\mathbf{g}; \hat{x}_T) := \mathbb{E}^{\mathbf{g}}\Big[c_T(X_T) \mid \Delta_T, \Lambda_T^{1:K} \Big] \\ &= \int_{\mathscr{X}_T} c_T(X_T)\, \Pi_T(\Delta_T, \Lambda_T^{1:K})(X_T, \hat{X}_T)\, dX_T, \end{aligned} \tag{B.1}$$

and so (34) holds with equality. Suppose that (34) holds for $t+1$. Then,

$$\begin{aligned} &J_t(\mathbf{g}; \hat{x}_{t:T}) = \mathbb{E}^{\mathbf{g}}\Big[\sum_{l=t}^{T-1} \Big[c_l(X_l, U_l^{1:K}) + \beta \cdot |X_{l+1} - \hat{X}_{l+1}|^2 \Big] \\ &+ c_T(X_T) \mid \Delta_t, \Lambda_t^{1:K} \Big] \\ &= \mathbb{E}^{\mathbf{g}}\Big[c_t(X_t, U_t^{1:K}) + \beta \cdot |X_{t+1} - \hat{X}_{t+1}|^2 \\ &+ \sum_{l=t+1}^{T-1} \Big[c_l(X_l, U_l^{1:K}) + \beta \cdot |X_{l+1} - \hat{X}_{l+1}|^2 \Big] \\ &+ c_T(X_T) \mid \Delta_t, \Lambda_t^{1:K} \Big] \\ &= \mathbb{E}^{\mathbf{g}}\Big[\mathbb{E}^{\mathbf{g}}\Big[c_t(X_t, U_t^{1:K}) + \beta \cdot |X_{t+1} - \hat{X}_{t+1}|^2 \\ &+ \sum_{l=t+1}^{T-1} \Big[c_l(X_l, U_l^{1:K}) + \beta \cdot |X_{l+1} - \hat{X}_{l+1}|^2 \Big] \\ &+ c_T(X_T) \mid \Delta_t, \Lambda_t^{1:K}, U_t^{1:K} \Big] \mid \Delta_t, \Lambda_t^{1:K} \Big] \\ &\geq \mathbb{E}^{\mathbf{g}}\Big[\mathbb{E}^{\mathbf{g}}\Big[c_t(X_t, U_t^{1:K}) + \beta \cdot |X_{t+1} - \hat{X}_{t+1}|^2 \\ &+ V_{t+1}\big(\phi_t\big[\Pi_t(\Delta_t, \Lambda_t^{1:K})(X_t, \hat{X}_t), Y_{t+1}^{1:K}, U_t^{1:K}\big]\big) \mid \\ &\Pi_t(\Delta_t, \Lambda_t^{1:K}), U_t^{1:K} \Big] \mid \Delta_t, \Lambda_t^{1:K} \Big] \\ &= \mathbb{E}^{\mathbf{g}}\Big[V_t\big(\Pi_t(\Delta_t, \Lambda_t^{1:K})(X_t, \hat{X}_t)\big) \mid \Delta_t, \Lambda_t^{1:K} \Big] \\ &= V_t\big(\Pi_t(\Delta_t, \Lambda_t^{1:K})(X_t, \hat{X}_t)\big), \end{aligned} \tag{B.2}$$

where, in the inequality, we used the hypothesis and, in the last equality, we used (33). Thus, (34) holds for all t.

(b) We prove the second part of the theorem by induction too. For $t = T$,

$$\begin{aligned} &J_T(\mathbf{g}; \hat{x}_T) := \mathbb{E}^{\mathbf{g}}\Big[c_T(X_T) \mid \Delta_T, \Lambda_T^{1:K} \Big] \\ &= \int_{\hat{\mathscr{X}}_T} c_T(X_T)\, \Pi_T(\Delta_T, \Lambda_T^{1:K})(X_T, \hat{X}_T)\, dX_T. \end{aligned} \tag{B.3}$$

Suppose that (33) holds for $t+1$. Then

$$\begin{aligned} &\inf_{u_t^{1:K} \in \prod_{k\in\mathcal{K}} \mathcal{U}_t^k} \mathbb{E}^{\mathbf{g}}\Big[\sum_{l=t}^{T-1} \Big[c_l(X_l, U_l^{1:K}) + \beta \cdot |X_{l+1} - \hat{X}_{l+1}|^2 \Big] \\ &+ c_T(X_T) \mid \Delta_t, \Lambda_t^{1:K} \Big] \\ &= \inf_{u_t^{1:K} \in \prod_{k\in\mathcal{K}} \mathcal{U}_t^k} \mathbb{E}^{\mathbf{g}}\Big[c_t(X_t, U_t^{1:K}) + \beta \cdot |X_{t+1} - \hat{X}_{t+1}|^2 \\ &+ \sum_{l=t+1}^{T-1} \Big[c_l(X_l, U_l^{1:K}) + \beta \cdot |X_{l+1} - \hat{X}_{l+1}|^2 \Big] \\ &+ c_T(X_T) \mid \Delta_t, \Lambda_t^{1:K} \Big] \\ &= \inf_{u_t^{1:K} \in \prod_{k\in\mathcal{K}} \mathcal{U}_t^k} \mathbb{E}^{\mathbf{g}}\Big[\mathbb{E}^{\mathbf{g}}\Big[c_t(X_t, U_t^{1:K}) + \beta \cdot |X_{t+1} - \hat{X}_{t+1}|^2 \\ &+ \sum_{l=t+1}^{T-1} \Big[c_l(X_l, U_l^{1:K}) + \beta \cdot |X_{l+1} - \hat{X}_{l+1}|^2 \Big] \\ &+ c_T(X_T) \mid \Delta_t, \Lambda_t^{1:K}, U_t^{1:K} \Big] \mid \Delta_t, \Lambda_t^{1:K} \Big] \\ &= \inf_{u_t^{1:K} \in \prod_{k\in\mathcal{K}} \mathcal{U}_t^k} \mathbb{E}^{\mathbf{g}}\Big[\mathbb{E}^{\mathbf{g}}\Big[c_t(X_t, U_t^{1:K}) + \beta \cdot |X_{t+1} - \hat{X}_{t+1}|^2 \\ &+ V_{t+1}\big(\phi_t\big[\Pi_t(\Delta_t, \Lambda_t^{1:K})(X_t, \hat{X}_t), Y_{t+1}^{1:K}, U_t^{1:K}\big]\big) \mid \Pi_t(\Delta_t, \\ &\Lambda_t^{1:K}), U_t^{1:K} \Big] \mid \Delta_t, \Lambda_t^{1:K} \Big] \\ &= \mathbb{E}^{\mathbf{g}}\Big[V_t\big(\Pi_t(\Delta_t, \Lambda_t^{1:K})(X_t, \hat{X}_t)\big) \mid \Delta_t, \Lambda_t^{1:K} \Big] \\ &= V_t\big(\Pi_t(\Delta_t, \Lambda_t^{1:K})(X_t, \hat{X}_t)\big), \end{aligned} \tag{B.5}$$

where, in the third equality, we used the hypothesis, and in the fourth equality, $u_t^{1:K}$ achieves the infimum. Thus, (33) holds for all t.

For $t = 0$, (34) yields $J_0(\mathbf{g}; \hat{x}_{0:T}) = V_0\big(\Pi_0(\Delta_0, \Lambda_0^{1:K})(X_0, \hat{X}_0)\big)$. Taking expectations

$$J(\mathbf{g}; \hat{x}_{0:T}) = \mathbb{E}^{\mathbf{g}}\Big[V_0\big(\Pi_0(\Delta_0, \Lambda_0^{1:K})(X_0, \hat{X}_0)\big) \Big]. \tag{B.6}$$

By (34), it follows that for any other $\mathbf{g}' \in \mathcal{G}$,

$$J(\mathbf{g}'; \hat{x}_{0:T}) \geq \mathbb{E}^{\mathbf{g}}\Big[V_0\big(\Pi_0(\Delta_0, \Lambda_0^{1:K})(X_0, \hat{X}_0)\big) \Big]. \tag{B.7}$$

Appendix C. Proof of Lemma 4

Obviously, for $t = T$,

$$\begin{aligned} &V_T\big(\rho\ \Pi_T(\Delta_T, \Lambda_T^{1:K})\big) \\ &= \int_{\mathscr{X}_T} \int_{\mathscr{X}_T} c_T(X_T)\ \rho\ \Pi_T(\Delta_T, \Lambda_T^{1:K})(X_T, \hat{X}_t)\, dX_T\, d\hat{X}_t \\ &= \rho\ V_T\big(\Pi_T(\Delta_T, \Lambda_T^{1:K})\big). \end{aligned} \tag{C.1}$$

For $t = 0, \ldots, T-1$, by assigning $\Pi_t = \rho\ \Pi_t$ [recall $p(X_t, \hat{X}_t \mid \Delta_t, \Lambda_t^{1:K}) = \Pi_t(\Delta_t, \Lambda_t^{1:K})$], (33) becomes

$$\begin{aligned} &V_t\big(\rho\ \Pi_t(\Delta_t, \Lambda_t^{1:K})\big) \\ &= \inf_{u_t^{1:K} \in \prod_{k\in\mathcal{K}} \mathcal{U}_t^k} \Big[\int_{\mathscr{X}_t} \int_{\mathscr{X}_t} c_t(X_t, U_t^{1:K})\ \rho \\ &\cdot \Pi_t(\Delta_t, \Lambda_t^{1:K})(X_t, \hat{X}_t)\, dX_t\, d\hat{X}_t \\ &+ \int_{\mathscr{Y}_{t+1}} \int_{\mathscr{X}_{t+1}} \int_{\mathscr{X}_{t+1}} \int_{\mathscr{X}_t} \int_{\mathscr{X}_t} V_{t+1}\big(\rho\ \Pi_{t+1}(\Delta_{t+1}, \Lambda_{t+1}^{1:K})\big) \end{aligned}$$

$$\cdot\, p(Y_{t+1}^{1:K} \mid X_{t+1})\, p(X_{t+1}, \hat{X}_{t+1} \mid X_t, \hat{X}_t, U_t^{1:K})\, \rho$$
$$\cdot p(X_t, \hat{X}_t \mid \Delta_t, \Lambda_t^{1:K}) dX_t\, d\hat{X}_t\, dX_{t+1}\, d\hat{X}_{t+1}\, dY_{t+1}^{1:K} \Big], \tag{C.2}$$

where $\mathcal{Y}_{t+1} = \otimes_{k\in\mathcal{K}} \mathcal{Y}^k$.

Next, from (31),

$$\rho\, \Pi_{t+1}(\Delta_{t+1}, \Lambda_{t+1}^{1:K})$$
$$= \frac{p(Y_{t+1}^{1:K} \mid X_{t+1}) \int_{\mathcal{X}_t} \int_{\mathcal{X}_t} p(X_{t+1}, \hat{X}_{t+1} \mid X_t, \hat{X}_t, U_t^{1:K}) \cdot \cancel{\rho} \cdot p(X_t, \hat{X}_t \mid \Delta_t, \Lambda_t^{1:K})\, dX_t\, d\hat{X}_t}{\int_{\mathcal{X}_{t+1}} \int_{\mathcal{X}_{t+1}} p(Y_{t+1}^{1:K} \mid X_{t+1}) \int_{\mathcal{X}_t} \int_{\mathcal{X}_t} p(X_{t+1}, \hat{X}_{t+1} \mid X_t, U_t^{1:K}) \cdot \cancel{\rho}\, p(X_t, \hat{X}_t \mid \Delta_t, \Lambda_t^{1:K})\, dX_t\, d\hat{X}_t\, dX_{t+1}},$$
$$= \Pi_{t+1}(\Delta_{t+1}, \Lambda_{t+1}^{1:K}). \tag{C.3}$$

Substituting (C.3) into (C.2), we have

$$V_t\big(\rho\, \Pi_t(\Delta_t, \Lambda_t^{1:K})\big)$$
$$= \inf_{u_t^{1:K} \in \prod_{k\in\mathcal{K}} \mathcal{U}_t^k} \Big[\int_{\mathcal{X}_t} \int_{\mathcal{X}_t} c_t(X_t, U_t^{1:K})\, \rho$$
$$\cdot\, \Pi_t(\Delta_t, \Lambda_t^{1:K})(X_t, \hat{X}_t)\, dX_t\, d\hat{X}_t$$
$$+ \int_{\mathcal{Y}_{t+1}} \int_{\mathcal{X}_{t+1}} \int_{\mathcal{X}_{t+1}} \int_{\mathcal{X}_t} \int_{\mathcal{X}_t} V_{t+1}\big(\Pi_{t+1}(\Delta_{t+1}, \Lambda_{t+1}^{1:K})\big)$$
$$\cdot\, p(Y_{t+1}^{1:K} \mid X_{t+1})\, p(X_{t+1}, \hat{X}_{t+1} \mid X_t, \hat{X}_t, U_t^{1:K})\, \rho$$
$$\cdot p(X_t, \hat{X}_t \mid \Delta_t, \Lambda_t^{1:K}) dX_t\, d\hat{X}_t\, dX_{t+1}\, d\hat{X}_{t+1}\, dY_{t+1}^{1:K} \Big]$$
$$= \rho\, V_t\big(\Pi_t(\Delta_t, \Lambda_t^{1:K})\big). \tag{C.4}$$

Appendix D. Proof of Theorem 4

Starting with (33), we have

$$V_t\big(\Pi_t(\Delta_t, \Lambda_t^{1:K})\big)$$
$$= \inf_{u_t^{1:K} \in \prod_{k\in\mathcal{K}} \mathcal{U}_t^k} \Big[\int_{\mathcal{X}_t} \int_{\mathcal{X}_t} c_t(X_t, U_t^{1:K}) \tag{D.1}$$
$$\cdot\, \Pi_t(\Delta_t, \Lambda_t^{1:K})(X_t, \hat{X}_t)\, dX_t\, d\hat{X}_t$$
$$+ \int_{\mathcal{Y}_{t+1}} \int_{\mathcal{X}_{t+1}} \int_{\mathcal{X}_{t+1}} \int_{\mathcal{X}_t} \int_{\mathcal{X}_t} V_{t+1}\big(\Pi_{t+1}(\Delta_{t+1}, \Lambda_{t+1}^{1:K})\big)$$
$$\cdot\, p(Y_{t+1}^{1:K} \mid X_{t+1})\, p(X_{t+1}, \hat{X}_{t+1} \mid X_t, \hat{X}_t, U_t^{1:K})$$
$$\cdot p(X_t, \hat{X}_t \mid \Delta_t, \Lambda_t^{1:K}) dX_t\, d\hat{X}_t\, dX_{t+1}\, d\hat{X}_{t+1}\, dY_{t+1}^{1:K} \Big], \tag{D.2}$$

where $\mathcal{Y}_{t+1} = \otimes_{k\in\mathcal{K}} \mathcal{Y}^k$.

Choosing

$$\rho = \int_{\mathcal{X}_{t+1}} \int_{\mathcal{X}_{t+1}} \int_{\mathcal{X}_t} \int_{\mathcal{X}_t} p(Y_{t+1}^{1:K} \mid X_{t+1})$$
$$\cdot\, p(X_{t+1}, \hat{X}_{t+1} \mid X_t, \hat{X}_t, U_t^{1:K}) \cdot p(X_t, \hat{X}_t \mid \Delta_t, \Lambda_t^{1:K})\, dX_t$$
$$\cdot\, d\hat{X}_t\, dX_{t+1}\, d\hat{X}_{t+1}, \tag{D.3}$$

we can use the positive homogeneity of $V_t\big(\Pi_t(\Delta_t, \Lambda_t^{1:K})\big)$ (Lemma 4) to write the second part of (D.2) as follows

$$\int_{\mathcal{Y}_{t+1}} \int_{\mathcal{X}_{t+1}} \int_{\mathcal{X}_{t+1}} \int_{\mathcal{X}_t} \int_{\mathcal{X}_t} V_{t+1}\big(\Pi_{t+1}(\Delta_{t+1}, \Lambda_{t+1}^{1:K})\big)$$
$$\cdot\, p(Y_{t+1}^{1:K} \mid X_{t+1})\, p(X_{t+1}, \hat{X}_{t+1} \mid X_t, \hat{X}_t, U_t^{1:K})$$
$$\cdot p(X_t, \hat{X}_t \mid \Delta_t, \Lambda_t^{1:K}) dX_t\, d\hat{X}_t\, dX_{t+1}\, d\hat{X}_{t+1}\, dY_{t+1}^{1:K}$$
$$= \int_{\mathcal{Y}_{t+1}} V_{t+1}\big(\rho\, \Pi_{t+1}(\Delta_{t+1}, \Lambda_{t+1}^{1:K})\big)\, dY_{t+1}^{1:K}$$
$$= \int_{\mathcal{Y}_{t+1}} V_{t+1}\Big(p(Y_{t+1}^{1:K} \mid X_{t+1}) \int_{\mathcal{X}_t} \int_{\mathcal{X}_t} p(X_{t+1},$$
$$\hat{X}_{t+1} \mid X_t, \hat{X}_t, U_t^{1:K}) \cdot p(X_t, \hat{X}_t \mid \Delta_t, \Lambda_t^{1:K})\, dX_t$$
$$d\hat{X}_t\Big)\, dY_{t+1}^{1:K}, \tag{D.4}$$

where, in the last equality, we substituted (D.3) and (A.8).

Thus, we can write (D.2) as

$$V_t\big(\Pi_t(\Delta_t, \Lambda_t^{1:K})\big)$$
$$= \inf_{u_t^{1:K} \in \prod_{k\in\mathcal{K}} \mathcal{U}_t^k} \Big[\int_{\mathcal{X}_t} \int_{\mathcal{X}_t} c_t(X_t, U_t^{1:K})$$
$$\cdot\, \Pi_t(\Delta_t, \Lambda_t^{1:K})(X_t, \hat{X}_t)\, dX_t\, d\hat{X}_t$$
$$+ \int_{\mathcal{Y}_{t+1}} V_{t+1}\Big(p(Y_{t+1}^{1:K} \mid X_{t+1}) \int_{\mathcal{X}_t} \int_{\mathcal{X}_t} p(X_{t+1},$$
$$\hat{X}_{t+1} \mid X_t, \hat{X}_t, U_t^{1:K}) \cdot p(X_t, \hat{X}_t \mid \Delta_t, \Lambda_t^{1:K})\, dX_t$$
$$d\hat{X}_t\Big)\, dY_{t+1}^{1:K} \Big]. \tag{D.5}$$

The remainder of the proof follows by induction. Suppose that $V_{t+1}\big(\Pi_{t+1}(\Delta_{t+1}, \Lambda_{t+1}^{1:K})\big)$ is concave. Since

$$V_{t+1}\Big(p(Y_{t+1}^{1:K} \mid X_{t+1}) \int_{\mathcal{X}_t} \int_{\mathcal{X}_t} p(X_{t+1},$$
$$\hat{X}_{t+1} \mid X_t, \hat{X}_t, U_t^{1:K}) \cdot p(X_t, \hat{X}_t \mid \Delta_t, \Lambda_t^{1:K})\, dX_t\, d\hat{X}_t\Big), \tag{D.6}$$

is the composition of a concave function and increasing linear function, it follows that it is concave. However, concavity is preserved by integration (see Boyd and Vandenberghe (2004), p. 79), hence

$$\int_{\mathcal{Y}_{t+1}} V_{t+1}\Big(p(Y_{t+1}^{1:K} \mid X_{t+1}) \int_{\mathcal{X}_t} \int_{\mathcal{X}_t} p(X_{t+1},$$
$$\hat{X}_{t+1} \mid X_t, \hat{X}_t, U_t^{1:K}) \cdot p(X_t, \hat{X}_t \mid \Delta_t, \Lambda_t^{1:K})\, dX_t$$
$$d\hat{X}_t\Big)\, dY_{t+1}^{1:K}. \tag{D.7}$$

is concave. Since the pointwise infimum of concave functions is concave, (D.5) is concave.

References

Aicardi, M., Davoli, F., & Minciardi, R. (1987). Decentralized optimal control of Markov chains with a common past information set. *IEEE Transactions on Automatic Control*, *32*(11), 1028–1031.

Akametalu, A. K., Fisac, J. F., Gillula, J. H., Kaynama, S., Zeilinger, M. N., & Tomlin, C. J. (2014). Reachability-based safe learning with Gaussian processes. In *53rd IEEE conference on decision and control* (pp. 1424–1431).

Armstrong, A. A., Johnson, A. J. W., & Alleyne, A. G. (2021). An improved approach to iterative learning control for uncertain systems. *IEEE Transactions on Control Systems Technology*, *29*(2), 546–555.

Arslan, G., & Yüksel, S. (2017). Decentralized Q-learning for stochastic teams and games. *IEEE Transactions on Automatic Control*, *62*(4), 1545–1558.

Åström, K., & Wittenmark, B. (1995). *Adaptive control*. Addison-Wesley Publising Company.

Aswani, A., Gonzalez, H., Sastry, S. S., & Tomlin, C. (2013). Provably safe and robust learning-based model predictive control. *Automatica*, *49*(5), 1216–1226.

Bertsekas, D. (2017). *Dynamic programming and optimal control* (4th ed.). Athena Scientific.

Bertsekas, D. P., & Tsitsiklis, J. N. (1996). *Neuro-dynamic programming*. Athena Scientific.

Bismut, J. (1973). An example of interaction between information and control: The transparency of a game. *IEEE Transactions on Automatic Control*, *18*(5), 518–522. http://dx.doi.org/10.1109/TAC.1973.1100388.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Brand, M. (1999). Structure learning in conditional probability models via an entropic prior and parameter extinction. *Neural Computation*, *11*(5), 1155–1182. http://dx.doi.org/10.1162/089976699300016395.

Chalaki, B., Beaver, L. E., Remer, B., Jang, K., Vinitsky, E., Bayen, A., & Malikopoulos, A. A. (2020). Zero-shot autonomous vehicle policy transfer: From simulation to real-world via adversarial learning. In *IEEE 16th international conference on control & automation* (pp. 35–40).

Dave, A., & Malikopoulos, A. A. (2019). Decentralized stochastic control in partially nested information structures. *IFAC-PapersOnLine*, *52*(20), 97–102.

Dave, A., & Malikopoulos, A. A. (2020). Structural results for decentralized stochastic control with a word-of-mouth communication. In *2020 American control conference* (pp. 2796–2801). IEEE.

Dydek, Z., Annaswamy, A., & Lavretsky, E. (2013). Adaptive control of quadrotor UAVs: A design trade study with flight evaluations. *IEEE Transactions on Control Systems Technology*, *21*, 1400–1406.

Eisen, M., Gatsis, K., Pappas, G. J., & Ribeiro, A. (2018). Learning in non-stationary wireless control systems via Newton's method. In *2018 annual American control conference* (pp. 1410–1417).

Fisac, J. F., Akametalu, A. K., Zeilinger, M. N., Kaynama, S., Gillula, J., & Tomlin, C. J. (2019). A general safety framework for learning-based control in uncertain robotic systems. *IEEE Transactions on Automatic Control*, *64*(7), 2737–2752.

Gatsis, K., & Pappas, G. J. (2021). Statistical learning for analysis of networked control systems over unknown channels. *Automatica*, *125*, Article 109386.

Guha, A., & Annaswamy, A. (2021). Online policies for real-time control using MRAC-RL. arXiv, arXiv:2103.16551.

Gupta, A., Yüksel, S., Başar, T., & Langbort, C. (2015). On the existence of optimal policies for a class of static and sequential dynamic teams. *SIAM Journal on Control and Optimization*, *53*(3), 1681–1712.

Gyorfi, L., & Kohler, M. (2007). Nonparametric estimation of conditional distributions. *IEEE Transactions on Information Theory*, *53*(5), 1872–1879. http://dx.doi.org/10.1109/TIT.2007.894631.

Howard, R. A. (1960). *Dynamic programming and Markov process*. The MIT Press.

Ioannou, P. A., & Sun, J. (1996). *Robust adaptive control*. PTR Prentice-Hall.

Kara, A. D., & Yüksel, S. (2018). Robustness to incorrect system models in stochastic control and application to data-driven learning. In *2018 IEEE conference on decision and control* (pp. 2753–2758). ISBN: 978-1-5386-1395-5, http://dx.doi.org/10.1109/CDC.2018.8619684.

Khong, S. Z., Nešić, D., & Krstić, M. (2016a). An extremum seeking approach to sampled-data iterative learning control of continuous-time nonlinear systems. *IFAC-PapersOnLine*, *49*(18), 962–967.

Khong, S. Z., Nešić, D., & Krstić, M. (2016b). Iterative learning control based on extremum seeking. *Automatica*, *66*, 238–245.

Kiumarsi, B., Vamvoudakis, K. G., Modares, H., & Lewis, F. L. (2018). Optimal and autonomous control using reinforcement learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(6), 2042–2062.

Krichene, W., Castillo, M. S., & Bayen, A. (2018). On social optimal routing under selfish learning. *IEEE Transactions on Control of Network Systems*, *5*(1), 479–488.

Krichene, W., Drighès, B., & Bayen, A. M. (2015). Online learning of Nash equilibria in congestion games. *SIAM Journal on Control and Optimization*, *53*(2), 1056–1081.

Krishnamurthy, V. (2016). *Partially observed Markov decision processes (From filtering to controlled sensing)*. Cambridge University Press.

Kumar, P. R., & Varaiya, P. (1986). *Stochastic systems: Estimation, identification and adaptive control*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., ISBN: 0-13-846684-X.

Kurtaran, B. (1979). Corrections and extensions to "decentralized stochastic control with delayed sharing information pattern". *IEEE Transactions on Automatic Control*, *24*(4), 656–657. http://dx.doi.org/10.1109/TAC.1979.1102080.

Kushner, H. J. (1971). *Introduction to stochastic control*. Holt, Rinehart and Winston.

Leman, T., Xargay, E., Dullerud, G., Hovakimyan, N., & Wendel, T. (2009). L1 adaptive control augmentation system for the X-48B aircraft. In *AIAA guidance, navigation, and control conference*.

Mahajan, A. (2008). *Sequential decomposition of sequential dynamic teams: Applications to real-time communication and networked control systems* (Ph.D. thesis), University of Michigan.

Mahajan, A., Martins, N. C., Rotkowitz, M. C., & Yüksel, S. (2012). Information structures in optimal decentralized control. In *2012 IEEE 51st IEEE conference on decision and control* (pp. 1291–1306). http://dx.doi.org/10.1109/CDC.2012.6425819.

Malikopoulos, A. A. (2009). Convergence properties of a computational learning model for unknown Markov chains. *Journal of Dynamic Systems, Measurement and Control*, *131*(4), 041011–041017.

Malikopoulos, A. A. (2016). A duality framework for stochastic optimal control of complex systems. *IEEE Transactions on Automatic Control*, *61*(10), 2756–2765.

Malikopoulos, A. A. (2023). On team decision problems with nonclassical information structures. *IEEE Transactions on Automatic Control*.

Malikopoulos, A. A., Papalambros, P. Y., & Assanis, D. N. (2010). Online identification and stochastic control for autonomous internal combustion engines. *Journal of Dynamic Systems, Measurement, and Control*, *132*(2), 024504.

McGuire, C. B., & Radner, R. (Eds.), (1972). *Decision and organization: A volume in honor of Jacob Marschak (studies in mathematical and managerial economics)*. North-Holland Pub. Co.

Narendra, K., & Annaswamy, A. (1989). *Stable adaptive systems*. Prentice-Hall, Inc.

Nayyar, A., Mahajan, A., & Teneketzis, D. (2011). Optimal control strategies in delayed sharing information structures. *IEEE Transactions on Automatic Control*, [ISSN: 00189286] *56*(7), 1606–1620. http://dx.doi.org/10.1109/TAC.2010.2089381.

Nayyar, A., Mahajan, A., & Teneketzis, D. (2013). Decentralized stochastic control with partial history sharing: A common information approach. *IEEE Transactions on Automatic Control*, *58*(7), 1644–1658.

Ooi, J. M., Verbout, S. M., Ludwig, J. T., & Wornell, G. W. (1997). A separation theorem for periodic sharing information patterns in decentralized control. *IEEE Transactions on Automatic Control*, *42*(11), 1546–1550. http://dx.doi.org/10.1109/9.649699.

Papadimitriou, C. H., & Tsitsiklis, J. (1982). On the complexity of designing distributed protocols. *Information and Control*, *53*(3), 211–218.

Papadimitriou, C. H., & Tsitsiklis, J. N. (1985). Intractable problems in control theory. In *1985 24th IEEE conference on decision and control* (pp. 1099–1103). http://dx.doi.org/10.1109/CDC.1985.268670.

Papadimitriou, C. H., & Tsitsiklis, J. N. (1987). The complexity of Markov decision processes. *Mathematics of Operations Research*, [ISSN: 0364-765X] *12*(3), 441–450.

Recht, B. (2019). A tour of reinforcement learning: The view from continuous control. *Annual Review of Control, Robotics, and Autonomous Systems*, *2*, 253–279.

Rosolia, U., & Borrelli, F. (2018). Learning model predictive control for iterative tasks. A data-driven control framework. *IEEE Transactions on Automatic Control*, *63*(7), 1883–1896.

Sahoo, P. P., & Vamvoudakis, K. G. (2020). On-off adversarially robust Q-learning. *IEEE Control Systems Letters*, *4*(3), 749–754.

Sastry, S., & Bodson, M. (1989). *Adaptive Control: stability, convergence and robustness*. Prentice-Hall, Inc..

Sondik, E. J. (1971). *The optimal control of partially observed Markov processes* (Ph.D. thesis), Stanford University.

Striebel, C. (1965). Sufficient statistics in the optimum control of stochastic systems. *Journal of Mathematical Analysis and Applications*, *12*(3), 576–592.

Subramanian, J., & Mahajan, A. (2019). Approximate information state for partially observed systems. In *2019 IEEE 58th conference on decision and control* (pp. 1629–1636).

Subramanian, J., Sinha, A., Seraj, R., & Mahajan, A. (2022). Approximate information state for approximate planning and reinforcement learning in partially observed systems. *Journal of Machine Learning Research*, *23*, 1–83.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Bradford Books.

Tsitsiklis, J., & Athans, M. (1985). On the complexity of decentralized decision making and detection problems. *IEEE Transactions on Automatic Control*, *30*(5), 440–446.

van Schuppen, J. H., & Villa, T. (2015). *Coordination control of distributed systems*. Springer.

Varaiya, P., & Walrand, J. (1978). On delayed sharing patterns. *IEEE Transactions on Automatic Control*, *23*(3), 443–445. http://dx.doi.org/10.1109/TAC.1978.1101739.

Witsenhausen, H. S. (1971). Separation of estimation and control for discrete time systems. *Proceedings of the IEEE*, [ISSN: 0018-9219] *59*(11), 1557–1566.

Witsenhausen, H. S. (1973). A standard form for sequential stochastic control. *Mathematical Systems Theory*, *7*(1), 5–11.

Wu, J., & Lall, S. (2014). A theory of sufficient statistics for teams. In *53rd IEEE conference on decision and control* (pp. 2628–2635). IEEE.

Wu, C., Parvate, K., Kheterpal, N., Dickstein, L., Mehta, A., Vinitsky, E., & Bayen, A. M. (2017). Framework for control and deep reinforcement learning in traffic. In *2017 IEEE 20th international conference on intelligent transportation systems* (pp. 1–8). ISBN: 978-1-5386-1526-3.

Zhai, L., & Vamvoudakis, K. G. (2021). A data-based private learning framework for enhanced security against replay attacks in cyber-physical systems. *International Journal of Robust and Nonlinear Control*, *31*(6), 1817–1833.

Andreas A. Malikopoulos received the Diploma in mechanical engineering from the National Technical University of Athens, Greece, in 2000. He received M.S. and Ph.D. degrees from the department of mechanical engineering at the University of Michigan, Ann Arbor, Michigan, USA, in 2004 and 2008, respectively. He is the Terri Connor Kelly and John Kelly Career Development Associate Professor in the Department of Mechanical Engineering at the University of Delaware, the Director of the Information and Decision Science (IDS) Laboratory, and the Director of the Sociotechnical Systems Center. Prior to these appointments, he was the Deputy Director and the Lead of the Sustainable Mobility Theme of the Urban Dynamics Institute at Oak Ridge National Laboratory, and a Senior Researcher with General Motors Global Research & Development. His research spans several fields, including analysis, optimization, and control of cyber–physical systems; decentralized systems; stochastic scheduling and resource allocation problems; and learning in complex systems. The emphasis is on applications related to smart cities, emerging mobility systems, and sociotechnical systems. He has been an Associate Editor of the IEEE Transactions on Intelligent Vehicles and IEEE Transactions on Intelligent Transportation Systems from 2017 through 2020. He is currently an Associate Editor of Automatica and IEEE Transactions on Automatic Control. He is a member of SIAM and AAAS. He is also a Senior Member of the IEEE and a Fellow of the ASME.